Simultaneous Linear Equations and the Determination of Eigenvalues

U. S. Department of Commerce
National Bureau of Standards
Applied Mathematics Series • 29

UNITED STATES DEPARTMENT OF COMMERCE • Sinclair Weeks, *Secretary*
NATIONAL BUREAU OF STANDARDS • A. V. Astin, *Director*

Simultaneous Linear Equations and the Determination of Eigenvalues

Edited by L. J. Paige and Olga Taussky

Proceedings of a symposium held August 23–25, 1951, in Los Angeles, California, under the sponsorship of the National Bureau of Standards, in cooperation with the Office of Naval Research

National Bureau of Standards
Applied Mathematics Series • 29
Issued August 31, 1953

For sale by the Superintendent of Documents, U. S. Government Printing Office, Washington 25, D. C. - - - - - - Price $1.50

Foreword

The importance of the study of the solution of systems of linear equations and the determination of eigenvalues has long been recognized. Many problems find their most direct and natural formulations as systems of linear equations. Then, too, often a system of linear equations furnishes the most convenient tool for an approximate analysis of a continuous mathematical model; classical cases are the approximate solution of differential equations by difference equations and the analogous discrete approach to integral equations. Similarly, the determination of eigenvalues of continuous or discrete vibrating systems is a fundamental problem in both classical and modern theoretical physics.

With the advent of the high speed automatic digital computing machines there came a compelling reason to take another look at these problems, and to study them in the light of the equipment now being made available. This reexamination has been going on for some ten years in various centers and a number of interesting new results have been announced. It therefore seemed desirable to organize a Symposium on these topics, where active workers could meet and exchange ideas. The generous support of the Office of Naval Research on the occasion of the National Bureau of Standards Semicentennial celebration made a comprehensive gathering possible. It took place at the National Bureau of Standards Institute for Numerical Analysis at Los Angeles on August 23–25, 1951.

The choice of location was particularly appropriate, for the mathematicians of the NBS Institute for Numerical Analysis have been devoting a considerable amount of effort to the topic of the Symposium. This specialization began under the leadership of Dr. J. Barkley Rosser in 1949–1950; its results are partially reflected in the reports presented at the Symposium by members of the National Bureau of Standards' Staff.

The technical organization of the meeting was mainly the responsibility of Dr. Olga Taussky-Todd, who, with the assistance of Dr. L. J. Paige, edited the present volume.

The Symposium consisted of invited reports from various mathematicians, on selected topics, followed by a round-table discussion. The reports on simultaneous linear equations were confined to finite systems, but those on the determination of eigenvalues dealt with both the discrete and continuous cases, and included reports specifically on the determination of bounds for the eigenvalues. The present volume contains the majority of the reports which were presented at the meeting.

A. V. Astin, *Director,*
National Bureau of Standards.

Contents

1. Tentative Classification of Methods and Bibliography on Solving Systems of Linear Equations[*]

By George E. Forsythe[†]

I. Introduction

The scope of this article is indicated in the introductions to sections II, IV, and especially V. In the same places will be found acknowledgments to some of the persons who have helped build the bibliography.

This article is a collection of working notes, and not a polished bibliography. Because there is apparently nothing comparable in the literature, it seems better to publish it as it is than to wait for a state of completion which might never arrive. Although many of the references in section V have been checked in the course of the preparation of the manuscript, time has not permitted a systematic editing and search for errors.[**] If readers will send in lists of errors and omissions, it may later be possible to issue a corrected edition of this article. Please address the compiler at the National Bureau of Standards, 405 Hilgard Avenue, Los Angeles 24, Calif.

II. Basis of the Tentative Classification; Symbols Used in the Outline

We are here dealing with methods for getting the solution $A^{-1}b$ of a system of linear equations $Ax=b$, where A is a nonsingular square matrix of order n and b is a column vector. We are equally concerned with methods for obtaining the inverse matrix A^{-1}; where such methods simply parallel those for getting $A^{-1}b$, they will not ordinarily be mentioned. Since the only known precedent for a classification of methods, JENSEN 1944,[1] covers but a fraction of the methods considered here, it was necessary to devise some basis for the outline of section III. The basis was developed after conversations with Dr. Theodore S. Motzkin, and is regarded as only tentative.

Consider, for the moment, an iterative process for getting $A^{-1}b$. Starting from b, A, and some initial vector $\xi^{(0)}$, a set of arithmetic operations is prescribed, which, if carried out without round-off or other error, will yield a new vector $\xi^{(1)}$: $\xi^{(1)}=F_1(b,A;\xi^{(0)})$. Similarly, $\xi^{(2)}$ is obtained by another set of operations, carried out without error. In general, $\xi^{(k)}$ is obtained from b, A, $\xi^{(k-1)}$, $\xi^{(k-2)}$, . . ., $\xi^{(0)}$ by applying a set of exact arithmetic operations which we assume to be uniquely described by a formula of type

$$\xi^{(k)}=F_k(b,A;\xi^{(k-1)},\xi^{(k-2)}, \dots, \xi^{(1)},\xi^{(0)}).$$

In this manner there is defined a theoretically determined sequence of vectors

$$\xi^{(0)},\xi^{(1)},\xi^{(2)}, \dots, \xi^{(k)}, \dots.$$

However, round-off errors occur in almost all numerical processes (there are exceptions). As a cumulative result of these errors, one actually obtains a vector $x^{(1)}$ instead of $\xi^{(1)}$. In the next step, one ordinarily uses b, A, and $x^{(1)}$ in the algorithm designed to compute $\xi^{(2)}$. Suppose that $\eta^{(2)}$ would be the result of errorless application of the algorithm:

$$\eta^{(2)}=F_2(b,A;x^{(1)},x^{(0)}),$$

[*] The preparation of this paper was sponsored (in part) by the Office of the Air Comptroller, USAF, and the Office of Naval Research.

[†] National Bureau of Standards, Los Angeles, Calif.

[**] In reading the galley proofs, the non-bracketed parts of the entries in section V have been corrected against the original papers or their reviews in MR, Zb, Fs, or SA (see section V), except for the perhaps 15 to 20 titles neither contained nor reviewed in the NBS Institute for Numerical Analysis.

[1] Names in capitals refer to the bibliography in section V.

where $x^{(0)}=\xi^{(0)}$. Because of round-off errors, one really obtains not $\eta^{(2)}$, but another uniquely defined vector $x^{(2)}$:

$$x^{(2)}=\hat{F}_2(b,A;x^{(1)},x^{(0)}).$$

Here $\hat{F}_2$ is the function F_2, altered in accord with each of the round-off errors entering the calculation.

In this manner will be determined a theoretical sequence of vectors

$$\eta^{(0)}(=\xi^{(0)}),\eta^{(1)},\eta^{(2)}, \ldots, \eta^{(k)}, \ldots,$$

and a practical sequence

$$x^{(0)}(=\xi^{(0)}),x^{(1)},x^{(2)}, \ldots, x^{(k)}, \ldots.$$

Each vector $x^{(k)}$ is determined from its precedessors by the "rounded-off function" $\hat{F}_k$:

$$x^{(k)}=\hat{F}_k(b,A;x^{(k-1)},x^{(k-2)}, \ldots,x^{(0)}).$$

Each vector $\eta^{(k)}$ is the result of errorless application of the theoretical function F_k to the "contaminated" sequence $x^{(0)}, \ldots, x^{(k-1)}$:

$$\eta^{(k)}=F_k(b,A;x^{(k-1)},x^{(k-2)}, \ldots,x^{(0)}).$$

We thus have three sequences in mind: (1) the sequence $\{\xi^{(k)}\}$ of vectors free from round-off error in all generations (it is these theoretical vectors with which convergence proofs generally deal), (2) the sequence $\{x^{(k)}\}$ of vectors which the computer actually obtains while making the prescribed round-off errors, and (3) the sequence $\{\eta^{(k)}\}$ of vectors which have one "generation" of errorless calculation, based on the computed vectors $x^{(0)}, \ldots, x^{(k-1)}$.

We have supposed heretofore that we deal with an iterative process, but the same analysis applies also to a direct process—for example, elimination. The result of an entire elimination solution is theoretically the vector $\xi^{(1)}=A^{-1}b$. In practice one makes round-off errors and obtains $x^{(1)}$ instead. Since $b-Ax^{(1)}=r^{(1)}\neq 0$, it is customary to solve the equation $A(\eta^{(2)}-x^{(1)})=r^{(1)}$ for the correction to be added to $x^{(1)}$ to obtain $\eta^{(2)}$. Here $\eta^{(2)}=A^{-1}b$, but round-off errors cause the approximate solution $x^{(2)}$ to be obtained instead, etc. Thus the earlier model describes also the "direct" solution of the system $Ax=b$ whenever round-off errors are made in the actual solution.

The present classification of methods is based on two trichotomies.

First trichotomy: The sequence $\{\eta^{(k)}\}$ is a priori bound to be of one of the following three types:

I. All $\eta^{(k)}$ are necessarily equal to $A^{-1}b$, i. e., all are exact.

II. Some $\eta^{(k)}$, but not all, are necessarily equal to $A^{-1}b$.

III. No $\eta^{(k)}$ is equal to $A^{-1}b$, except for special choices of $\eta^{(0)}$, b, A.

Typical processes of these three types are I, the elimination process of Gauss 1826; II, the finite iteration of Lanczos 1951; and III, the infinite iteration of Seidel 1874.

Second trichotomy: The sequence $\{F_k\}$ of functions is bound a priori to fall into one of the following three classes:

α. All functions F_k have in fact the same number of arguments and are identical, i. e., there exists an integer, r, such that

$$(*) \qquad F_k(b,A;x^{(k-1)},\ldots,x^{(0)})\equiv F(b,A;x^{(k-1)},\ldots,x^{(k-r)}) \text{ (all } k).$$

Such processes are called *stationary*.

β. There are a finite number (greater than one) of distinct functions F_k, each of type (*). Such processes are called *partly stationary*.

γ. There are an infinite number of distinct functions F_k. Such processes are called *nonstationary*.

Typical processes of these types are α, (iterated) elimination; β, iterated elimination, occasionally interrupted by an acceleration procedure; γ, weighted averages of the successive iterates of Seidel's process, with weights equal to the coefficients of the Chebyshev polynomials of order 1, 2, 3, ... (Divisions between processes of types β and γ could be made in many different ways.)

The double trichotomy gives us a priori a nine-way classification of methods for getting $A^{-1}b$ (or indeed of estimating any quantity by a numerical process). Three of the possible types, Iβ, Iγ, and

IIβ, do not seem to occur in the literature on getting $A^{-1}b$, but the other six form a convenient structure on which to hang an outline of methods of solving linear equations:

	α Stationary process	β Partly stationary process	γ Nonstationary process
I. All $\eta^{(k)}$ exact	Iα	----------	----------
II. Some $\eta^{(k)}$ exact	IIα	IIβ	----------
III. No $\eta^{(k)}$ exact	IIIα	IIIβ	IIIγ

It is not presumed that this particular nine-way classification is necessarily a "good" one; it merely provides one of many possible frameworks on which to spread out the known solution processes.

The outline in section III serves two purposes: (1) It contains a tentative classification of methods for obtaining A^{-1} or $A^{-1}b$; (2) it provides subject headings for the bibliography of section V. Purpose (1) is achieved by the portion of the outline in which the six major headings Iα, IIα, IIβ, IIIα, IIIβ, IIIγ are subdivided. In these the methods are classified according to the nature of the iteration functions found in the literature. No strenuous effort has been made to bring all closely related methods into a single entry, even where this could be done within the conventional outline structure, for it is felt that purpose (2) is better served without too great a condensation. There are some methods that belong at two distinct points of the outline (e. g., Seidel's at 10c and at 13a (*1* & *5*)).

Purpose (2) requires not only the above subdivisions, but also the heading 0 (surveys), and those under IV (miscellaneous relevant topics) and V (contents unknown).

A good many words could be said in explanation of the outline and its notation, but it is hoped that the reader will be able to make it out well enough. Let it suffice to define the symbols used:

$A=(a_{ij})$=nonsingular square matrix of order n;
b=column vector;
$f(A)$=characteristic polynomial of A;
I=unit matrix;
A^{-1}=inverse of A;
A^T=transpose of A;
"$A>0$" means A is positive definite;
δ^2 process: see AITKEN 1925, 1937b, 1950;
$x^{(k)}$=kth approximating vector to $A^{-1}b$;
B=matrix approximating to A which is effectively inverted in processes of type 10;
$|x|_R=(x^TRx)^{\frac{1}{2}}$=length of x in R metric;
e_i=unit column vector with 1 for ith component and 0 for all other components;
$X^{(k)}$ is kth approximating matrix to A^{-1};
d_i=some column vector;
$\Delta x^{(k)}=x^{(k+1)}-x^{(k)}$;
P, Q=nonsingular square matrices of order n.

III. Tentative Outline of Methods

0. Survey of methods.
   a. Mainly iterative methods.
   b. Mainly direct methods.
   c. Contains references not included in section V.

Iα. *All answers theoretically exact. Stationary process.*

1. Solve explicitly.
   a. By determinants.
      1. Efficient evaluation of determinants (often same as 2a).
   b. Otherwise.

2. Triangularize A and solve.

a. By elimination (almost same as 3a*1*).

1. Inexact.

A. Efficient arrangements.

2. Exact.

A. Controlled magnitude of numbers.

b. By extension≡escalator processes (related to 4a).

1. By A-orthogonalization of unit vectors.

c. By generalized Cholesky method (A symmetrical or not).

1. Square-root method (symmetric A).

d. Any of above, with search for pivots.

e. Any of above, with "below-the-line" arrangement.

f. Any of above, using Cracovians.

3. Orthogonalize A and solve.

a. By triangularization.

1. Elimination (almost same as 2a).

2. Complete elimination.

3. Above, with search for pivots.

b. By orthogonalizing rows or columns of A.

4. Partition A into blocks, and carry out processes 2 or 3 on the blocks.

a. Smaller matrices inverted by a direct process.

b. Smaller matrices inverted by iterative process.

5. Iterate any of above.

IIα. *Some answers theoretically exact. Stationary process.*

6. Methods related to characteristic polynomial f of some matrix H.

a. Calculate $f(A)$ from determinant $d(A-\lambda I)$.

1. Use Newton's identities.

b. Finite iteration to get $f(A)$.

1. Matrix iteration.

2. Vector iteration.

c. Process 6b*2* applied to iterates of an infinite iteration, to get $f(H)$.

d. Methods related to successively orthogonalizing the vectors Ax, A^2x, A^3x, A^4x, . . .

1. In I metric.

2. In A^{-1} metric (when $A=A^T>0$).

7. Not related to characteristic polynomials.

a. Use of theory of congruences in n dimensions.

b. Replacement methods (closely related to 2).

1. Simplex method.

2. Replace one column of A at a time.

IIβ. *Some answers theoretically exact. Partly stationary process.*

8. Occasionally interrupt the iteration of process of type Iα by accelerating with . . .

a. . . . δ^2 process (21b*1*).

b. . . . linear combinations of prior answers (i. e., apply 21a*1*).

c. . . . other process.

9. Occasionally interrupt the iterative cycles of process of type IIα by accelerating . . .

a. . . . with δ^2 process (21b*1*).

b. . . . with linear combinations of prior answers (i. e., apply 21a*1*).

c. . . . by starting over from current residual (for process 6d).

d. . . . with other process.

IIIα. *No answer theoretically exact. Stationary process.*

10. Use a linear iteration function: $x^{(k+1)}=x^{(k)}+B^{-1}(b-Ax^{(k)})$.

a. General process of first degree.

b. B=diagonal of A (Jacobi).

c. B=lower triangle of A (Seidel).*

d. B=scalar matrix (e. g., Neumann series).

e. B built of submatrix blocks.

1. Blocks inverted by type I process.

2. Blocks inverted by type II process.

f. Reflection in hyperplanes.

g. Linear processes of second or higher degree (actually 21a*1*).

11. Use a nonlinear iteration function.

a. Use a least-squares process with hermitian matric R (here $\pi=\pi(x)$ determines the subspace in which one minimizes $|Ax-b|_R$.).

1. $\pi=Ax-b$ (one dimension).

2. $\pi=A^T(Ax-b)$ (one dimension).

3. π has 2 or more dimensions.

4. Systematically undershoot or overshoot point of minimum $|Ax-b|_R$.

5. $R=I$.

6. $R=A^{-1}$ (when $A=A^T>0$).

7. $R=(AA^T)^{-1}$.

8. Other R.

b. Use nonhermitian metric or nonmetric process.

1. Use piecewise linear iteration (relaxation methods): For some i maximizing . . .

(choice of one)

A. . . . $|(Ax-b)_i|$, change . . .

B. . . . $|(Ax-b)_i|/|a_{ii}|$, change . . .

C. . . . $|(Ax-b)_i|/|a_{ii}|^{\frac{1}{2}}$, change . . .

D. . . . $|[A^T(Ax-b)]_i|/(A^TA)_{ii}^{\frac{1}{2}}$, change . . .

E. . . . something else, change . . .

(choice of one)

F. . . . e_i (unit vector) so as to bring . . .

G. . . . A^Te_i so as to bring . . .

H. . . . something else so as to bring . . .

(choice of one)

I. . . . $(Ax-b)_i$ approximately to 0.

J. . . . $[A^T(Ax-b)]_i$ approximately to 0.

K. . . . something else to 0.

L. With any scheme above, systematically overshoot or undershoot point where residual becomes 0 (over and underrelaxing).

2. Use piecewise linear iteration (block relaxation methods): For some two or more i, change two or more . . .

(choice of one)

A. . . . e_i so as to minimize something.

B. . . . A^Te_i so as to minimize something.

c. Get A^{-1} with polynomial iteration function:

1. Use $X^{(k+1)}=X^{(k)}\ (2I-AX^{(k)})$.

2. Modify *1*.

3. Use related polynomials of third or higher order.

IIIβ. *No answer theoretically exact. Partly stationary process.*

12. Occasionally interrupt the iteration of process of type IIIα by accelerating with . . .

a. . . . δ^2 process (21b*1*).

b. . . . step of type 11a*4* (in types 11a).

c. . . . change of π.

d. . . . step of type 11b*2* (in types 11b*1*).

e. . . . change of B (in type 10).

f. . . . linear combination of prior answers (i. e., apply 21a*1*).

*Added in proof: This is commonly called Seidel's process. Professor Ostrowski has pointed out, however, that SEIDEL 1874 advocated only the process 11b*1* (*C & F & I*), while the process 10c was apparently first advocated by NEKRASOV 1884 (see Second Supplement), who attributed it to Seidel.

13. Use m iteration functions in cyclic order (when consolidated in cycles, process becomes typ IIIα.)

a. Use least-squares process with hermitian metric R, minimizing along directions d_1, d_2, ... , d_n, and repeat.

1. $d_i = e_i$ (unit vectors).

2. $d_i = A^T e_i$.

3. Systematically undershoot or overshoot point of minimum $|Ax - b|_R$.

4. $R = I$.

5. $R = A^{-1}$ (when $A = A^T > 0$, 13a (*1* & *5*) = Seidel process 10c).

6. $R = (AA^T)^{-1}$.

7. Other R.

b. Use least-squares process with hermitian metric R, minimizing along directions d_1, d_2, . . ., d_n, d_{n-1}, . . ., d_2, and repeat.

1. $d_i = e_i$ (unit vectors), $R = A^{-1}$ (back-and-forth Seidel process).

2. $d_i = e_i$, general R.

c. Use m linear iteration functions of type 10a and repeat.

1. Use them in a "back-and-forth" order.

d. Use m nonlinear iteration functions.

14. Use one of n separate processes of type 10a in arbitrary order.

a. Any process of type 11b*1* can be regarded as a repeated choice among n processes o type 10a.

b. Use some permutation of order 1, 2, . . ., n at each cycle.

IIIγ. *No answer theoretically exact. Nonstationary process.*

15. Use infinitely many distinct iteration functions.

a. Use polynomials in A of orders 1, 2, 3, ...

1. Generalized Chebyshev polynomials.

2. Other polynomials.

b. Polynomials in other matrices connected with iterative processes.

c. Monte Carlo method (estimates certain population means related to A^{-1}, by samples of size 1, 2, 3, . . .).

1. Based on Neumann series 10d.

IV. *Miscellaneous relevant topics.*

16. Analyses of errors.

a. Not related to specific method.

1. "Condition".

2. Stability of solution, i. e., variation of A^{-1} (or $A^{-1}b$), as A varies (or A and b vary).

b. Related to specific method.

1. Truncation errors.

2. Round-off errors.

A. Worst possible.

B. Probabilistic.

3. Control of blunders.

17. "Preconditioning" (i.e., changing the system $Ax = b$).

a. By enlarging A.

1. By Gauss' transformation.

2. By homogenizing the system (special case of *1*).

b. By premultiplication of A.

1. By A^T (normalization).

2. By some polynomial in A.

3. By some approximation to A^{-1}.

c. By approximate use of some method of solution.

1. Orthogonalization.

2. Elimination.

d. By taking high power of A (after 17a*2*).
e. By successively rotating off-diagonal elements of A into diagonal (when $A=A^T$).
18. Methods for punched-card equipment.
19. Methods for automatic digital computing machinery.
19A. Methods for analogue computing machinery.
20. Special computing tricks.
a. Cumulating solution as $\Sigma\Delta x^{(k)}$.
1. Using few significant digits at once.
b. Getting A^{-1} by solving n systems $Ax=e_i$ simultaneously (operating on I).
c. Getting $A^{-1}B$ by applying to B the row operations that reduce A to I.
d. If $PAQ=I$, then $A^{-1}=QP$.
e. Splitting real parts of A from imaginary parts of A.
21. Methods for accelerating convergence, applicable to many problems of numerical analysis.
a. Linear processes.
1. Use of linear combinations of sequence of $x^{(k)}$.
b. Nonlinear processes.
1. Use of δ^2 process.
2. Use of generalized δ^2 process.
22. Counts of operations, time, labor, etc.
23. Methods applicable to special matrices.
a. Almost triangular matrices.
b. Matrices of large order.
V. *Contents or part of contents unknown.*

IV. Cross References From the Outline to the Bibliography

In this section the subject heading (numbers and letters like 11e*1*) correspond to the outline in section III. Opposite each subject heading are given the references from the bibliography (section V) that deal with the subject. There is an inconsistency in the presentation about which the user should be warned. For example, Kunz 1947 treats several of the methods 2a, 2b, 2c, . . . , but is listed only under the general heading 2, whereas Bodewig 1947b will be found under 2a, 2b, and 2c. (The unavailability of Kunz 1947 at the time of compiling section IV caused the difference in treatment.)

A question mark after an entry indicates doubt as to whether the title deals with the subject, and does not query the date as such.

0. Bargmann-Montgomery-von Neumann 1946, Bodewig 1947b, Engineering Research Associates 1950, Forsythe 1951, Fox 1950a, Frazer-Duncan-Collar 1938, Grossman 1950?, Hartree 1949b, Householder 1949, 1950, Kunz 1947, Pipes 1948, Salvadori 1948.

a. Geiringer 1949, Hotelling 1943a, 1949, Panov 1934, Reich 1948, Zurmühl 1950.

b. Dwyer 1951, Jensen 1944, Kunz 1947, Turing 1948, Zurmühl 1950.

c. Dunlap 1939, Dwyer 1941a, 1942, 1951, Frame 1945, Griffin 1934, Higgins 1949, Jenne 1949, Kantorovich-Krylov 1948, International Business Machines Corp. 1950, Southwell 1940, 1946, Taussky 1951, Tolley-Ezekiel 1927, zur Capellen 1948.

1α. Adams 1945?

1.

a. Bargmann-Montgomery-von Neumann 1946, Bodewig 1947a, Cassina 1948, Dickson 1922, Ingraham 1937, Leppert 1947, Milne 1949, Ostrowski 1938, Salvadori 1948, Schmeidler 1949, H. Schulz 1938.

1. Aitken 1937a, Chiò 1853, Collatz 1948a, Dwyer 1951, Goodwin 1950, Jensen 1944, Milne-Thomson 1941, Pipes 1948, Rice 1920, Smith 1927, Walther 1944, Whittaker-Robinson 1924.

b. Bisshopp 1945, Collar 1939, Jenne 1949, Nörlund 1940.

2. Kunz 1947.

a. Aitken 1932, 1935, Andersen 1947, Andree 1951, Baetslé 1951, Bargmann-Montgomery-von Neumann 1946, Bauschinger 1901, Bodewig 1947b, Cassina 1948, Cassinis 1946, Collatz 1948a, Couffignal 1944, Deming 1928, Dwyer 1941a, 1951, Encke 1835, Forsythe 1951, Fox 1950a, Fox-Huskey-Wilkinson 1948a, Frazer-Duncan-Collar 1938, Fröberg 1953, Gauss 1826, Harvard Computation Lab. 1946, Herzberger 1949, Herzberger-Morris 1947, Hoel 1941, Householder 1949, 1950, Jensen 1944, 1948, Jürgens 1886, Kolmogoroff 1946?, Margenau-Murphy 1943, Mehmke 1930, Milne 1949, Mitchell 1948, Parkes 1950, Polachek 1948, Renner 1946, Reynolds 1934, Ricci 1949, D. F. Richardson 1946, Salvadori 1948, Scarborough 1950, H. Schulz 1938, Seidel 1874, Spoerl 1943, 1944, Verzuh 1949, Walther 1944, Whittaker-Robinson 1924, Willers 1928, 1947, Worch 1932, Wren 1937, Zurmühl 1950.

1. Fox 1950a, Goldstine-von Neumann 1951, Pipes 1948, Turing 1948.

A. Black 1949, Bruner 1947, Doolittle 1878, Dwyer 1941d, 1951, Fox 1950a, Kurtz 1936, Leavens 1947, Nielsen-Goldstein 1947, Satterthwaite 1944, Tollay-Ezekiel 1927, Waugh-Dwyer 1945, Rosser 1952.

2. Dodgson 1866, Dwyer 1951, Waugh-Dwyer 1945, Rosser 1952.

b. Aitken 1932, Bodewig 1947b, Boschan 1946, Cochran 1938?, Collatz 1948a, Dwyer 1951, Frazer 1947, Householder 1949, 1950, Jossa 1938?, 1940?, Morris 1946, 1947, Neville 1948, Turing 1948.

1. Fox 1950a, Fox-Huskey-Wilkinson 1948a, Turing 1948.

c. Andersen 1947, Baetslé 1951, Banachiewicz 1951, Bodewig 1947b, 1950, Cassina 1948, Cassinis 1946, Crout 1941, Dwyer 1951, Forsythe 1951, Fox 1950a, 1950b, Fox-Huskey-Wilkinson 1948a, Herzberger 1949, Householder 1949, 1950, Istituto per le Applicazioni del Calcolo #4 (date unknown), Jensen 1944, Milne 1949, Picone (date unknown), Ricci 1949, Turing 1948, Waugh-Dwyer 1945, Zurmühl 1949, 1950.

1. Asplund 1945, Banachiewicz 1938a?, 1938b, Benoit 1924, D. B. Duncan-Kenney 1946, Dwyer 1942, 1944, 1945, Istituto per le Applicazioni del Calcolo #3 (date unknown), Jensen 1944, Kunz 1947, Laderman 1948, Rubin 1926.

d. Dwyer 1951, Fox 1950a, Fox-Huskey-Wilkinson 1948a, Frazer-Duncan-Collar 1938, Herzberger 1949, Herzberger-Morris 1947, Waugh-Dwyer 1945, Zurmühl 1950.

e. Andree 1951, Baetslé 1951, Dwyer 1951, Forsythe 1951, Fox 1950a, Hoel 1941, Verzuh 1949.

f. Backman 1946, Banachiewicz 1937b, 1937c, 1942, 1951, Kamela 1943.

3.

a. Basile 1949, Kunz 1947?

1. Bodewig 1947b, Fox-Huskey-Wilkinson 1948a, Frazer-Duncan-Collar 1938, Jürgens 1888.

2. Clasen 1888, Dreyer 1943, Forsythe 1951, Fox 1950a, Householder 1949, Jensen 1944, Jordan 1920, Petrie 1953, Samuelson 1950, Turing 1948, Volta 1950, Whittaker-Robinson 1924, Worch 1932.

3. Forsythe 1951.

b. Bargmann-Montgomery-von Neumann 1946, Cassinis 1946, Forsythe 1951, Fox 1950a, Householder 1950, Kunz 1947?, Picone (date unknown), Ricci 1949, Roma 1946, 1947, E. Schmidt 1908, Shreĭder 1951, Turing 1948, Unger 1951.

4. Banachiewicz 1937a, 1937b, 1937c?, Bargmann-Montgomery-von Neumann 1946, Bodewig 1947a, 1947b, Boltz 1923, Collatz 1948a, W. J. Duncan 1944, Forsythe 1951, Fox 1950a, Frazer-Duncan-Collar 1938, (Gram 1883), Guttman 1946, Hotelling 1943a, 1943b, 1949, Ingraham 1937, Jossa 1938, 1940, Jensen 1944, Krüger 1905, Kunz 1947?, Morris 1946, 1947, Ostrowski 1938, Pipes 1941, Saibel 1944, Schur 1917, Turetsky 1951, Waugh 1945, Woodbury 1950.

a. (Many of references under 4.)

b. Neville 1948.

5. Bodewig 1947b, Forsythe 1951, Polachek 1948, Runge-König 1924, Salvadori 1948, **Thomson** 1878, Turing 1948, Zurmühl 1944, 1950.

IIα.

6.

a. Dwyer 1951.

1. Hotelling 1943a.

b.

1. Bingham 1941, Dwyer 1951, Frame 1949, Pipes 1948.

2. Forsythe 1951, Freeman 1943, Milne 1951.

c. Freeman 1943, R. J. Schmidt 1941.

d. Hestenes 1951, Lanczos 1951, Stiefel 1951, 1952, Rosser 1953.

7.

a. Lewy (unpublished), Robinson 1951.

b.

1. Dantzig 1951.

2. Sherman 1951, Sherman-Morrison 1949.

IIβ.

8.

9.

c. Hestenes (unpublished).

IIIα.

10. Argelander 1844, Bauschinger 1901, Cherepkov 1936, Grossman 1948, Kantorovich 1939, Kunz 1947, Waddell 1916, Zylev 1939.

a. Aitken 1950, Cesari 1937a, 1937b, Collatz 1950b, Frazer-Duncan-Collar 1938, Forsythe 1951, Geiringer 1949, Hotelling 1943, Householder 1949, Milne 1951, Morris 1935?, Pipes 1948, Reich 1948, Wittmeyer 1936b.

b. Anér 1926, Black (date unknown), Brand 1935?, Cassinis 1946, Chen 1944?, Collatz 1942, 1948a, 1949, 1950a, 1950b, Cross 1932, Forsythe 1951, Fox 1950a, Frazer-Duncan-Collar 1938, Geiringer 1949, Householder 1949, 1950, Ivanov 1939, Jacobi 1845, Kormes 1943?, 1947?, Morris 1947, Runge-König 1924, Schott 1855, Stein-Rosenberg 1948, von Mises Pollaczek-Geiringer 1929, Young 1950.

c. Aitken 1950, Bowie 1950, Cassinis 1946, Collatz 1942, 1948a, 1949, 1950a, 1950b, Forsythe 1951, Fox 1950a, Frankel 1950, Frazer-Duncan-Collar 1938, Gatto 1949, Geiringer 1949, Hotelling 1933, 1943a, Householder 1949, 1950, Ivanov 1939, Kantorovich-Krylov 1948, Liebmann 1918, Mehmke 1892, Mehmke-Nekrassof 1892?, Miller 1947, Morris 1935, 1947, Oldenburger 1940b, Pipes 1948, Pollaczek-Geiringer 1928, Reich 1948, 1949, Runge 1899, Salvadori 1948, Sassenfeld 1951, Schmeidler 1949, Seidel 1862, 1874,* Shortley-Weller 1938, Shortley-Weller-Fried 1940, Snyder-Livingston 1949, P. Stein 1951, Stein-Rosenberg 1948, von Mises-Pollaczek-Geiringer 1929, Whittaker-Robinson 1924, Willers 1928, 1947, Young 1950, Zurmühl 1950.

d. Biezeno 1924, Bodewig 1947b, Forsythe 1951, Frankel 1950, Geiringer 1949, Hellinger-Toeplitz 1924, Holley 1951, Hotelling 1943a, Householder 1949, Milne 1951, Newing 1941, Pipes 1948, Plunkett 1950, Quade 1947, Richardson 1910, Schmeidler 1949, von Mises-Pollaczek-Geiringer 1929, Walsh 1920, Waugh 1950.

e.

1. Gauss 1826, Geiringer 1942, 1949, Hertwig 1912, Mehmke 1892.

f. Bodewig 1947b, Cimmino 1938, Forsythe 1951, Householder 1949.

g. Frankel 1950.

11.

a (comprehensive).

Forsythe 1951, Householder 1949, Rosser 1949, 1950, M. Stein 1952.

(*1* & *6*). Birman 1950?, Householder 1950, Kantorovich 1945, 1947, Mysovskikh 1950?, Pipes 1948, Temple 1939.

(*2* & *5*). Booth 1949, Cauchy 1847, Curry 1944, Forsythe-Motzkin 1950, 1951a, 1951b, Hartree 1948, 1949b, Kantorovich 1948, Kantorovich-Krylov 1948, Reich 1948.

(*3* & *5*). Forsythe-Motzkin 1950, 1951b.

(*3* & *6*). Kantorovich 1947.

(*2* & *4* & *5*). Cauchy 1847, M. Stein 1952.

4. Hartree 1948?, M. Stein 1952.

b.

1. Black-Southwell 1938, Bowie 1947?, Fox 1947, Hendrikz (date unknown)?, Menzies (date unknown), Nikolaeva 1949, L. Wright 1943.

(*A* & *F* & *I*). Black 1938, Black (date unknown), Forsythe 1951, Fox 1948, 1950b, Gaskell 1943, Householder 1950, Milne 1951, Reich 1948, Southwell 1940, 1946.

(*D* & *G* & *J*). Householder 1949.

(*B* & *F* & *I*). Dedekind 1901, Fox 1948, Gauss 1823, Gerling 1843, Jürgens 1886, Rainsford (date unknown), Schaefer 1927, Schott 1855, Southwell 1940, 1946, Temple 1939, Thompson (date unknown), Whittaker-Robinson 1924, Zurmühl 1950.

L. Fox 1948, Gerling 1843, Southwell 1940, 1946.

(*C* & *F* & *I*). Fox 1950a, Householder 1949, Seidel 1874, Synge 1944.

2. Black (date unknown), Fox 1948, 1950b, Southwell 1940, 1946.

*See footnote, p. 5.

c.

1. Bargmann-Montgomery-von Neumann 1946, Bodewig 1947b, Fox 1950a, Frazer-Duncan-Collar 1938, Hotelling 1943a, 1949, Householder 1949, Kunz 1947?, Lonseth 1949, Kantorovich 1949, Neville 1948, Pipes 1948, G. Schulz 1933, Turing 1948, Ullman 1944.

2. Bargmann-Montgomery-von Neumann 1946.

3. Bodewig 1947b, Schröder 1870, Ullman 1944.

IIβ.

12. Li͡usternik 1947, Shortley-Weller 1938, Shortley-Weller-Fried 1940.

a. Forsythe 1951.

b. Hestenes (unpublished).

c. Forsythe-Motzkin 1951b.

d. (References under 11b*2*.)

e. Forsythe 1951, Richardson 1910.

f. Fox 1950b, Milne 1951.

13.

a (comprehensive). Householder 1949.

(*1* & *4*). de la Garza 1951, Householder 1949, Rosser 1949.

(*2* & *6*). Bodewig 1947b, Bottema 1950, Kaczmarz 1937, Rosser 1949, Tompkins 1949.

(*1* & *5*). Frankel 1950, Householder 1950.

(*1* & *3* & *5*). Frankel 1950.

(*2* & *4*). Householder 1949.

b.

1. Aitken 1950.

c. Bückner 1950, Geiringer 1949, Richardson 1910.

1. Aitken 1950, Rosser 1949, 1950.

d. Kelley-Salisbury 1926.

14.

a. Hestenes (unpublished).

b. Geiringer 1949, Reich 1948.

IIIγ

15.

a. Flanders-Shortley 1950.

b.

c.

1. Forsythe-Leibler 1950, Opler 1951, Swift-Tikson 1951, Todd 1951a, 1951b, Wasow 1952.

IV.

16. Dwyer 1953, Opitz-Willers (date unknown), Pirlet 1909, H. Schulz 1938.

a. Hertwig 1905?, Price 1951, Sherman-Morrison 1950.

1. Hartree 1948, Jürgens 1888, Taussky 1949, 1950, Todd 1949a, 1949b, Turing 1948.

2. Bargmann-Montgomery-von Neumann 1946, Berkson 1936, Bartlett 1951, Blumenthal 1914, Collatz 1949, W. E. Deming 1937, Dwyer 1951, Etherington 1932, Forsythe 1951, Hotelling 1943a, Janet 1920a, 1920b, Lonseth 1942, 1944, 1947, Milne 1949, Morgenstern-Woodbury 1950, Moulton 1913, 1936, Ostrowski 1937a, 1937b, 1950, Redheffer 1948, Roessler 1936, Scarborough 1950, Tuckerman 1941, Turing 1948, Walsh 1920, Willers 1928, 1947, Wittmeyer 1934, 1936a, Woodbury 1949, 1950, Zurmühl 1950.

b. Hotelling 1949, Householder 1949, Leavens 1947.

1. Collatz 1950b, Wittmeyer 1936b.

2. Bargmann-Montgomery-von Neumann 1946, Dwyer 1951, Goldstine-von Neumann 1951, Neville 1948, Parkes 1950, Polachek 1948, Reich (date unknown), Reich 1948, Salvadori 1948, Satterthwaite 1944, Scarborough 1950, Tuckerman 1941, Turing 1948, Waugh 1950.

A. Hoel 1940, Hotelling 1943a, 1943b.

B. Hoel 1940, Hotelling 1943a, Ullman 1944.

3. Dwyer 1951, Gauss 1823.

17.

a.

1. Forsythe-Motzkin 1952, Gauss 1823, Gerling 1843, Zurmühl 1950.

b.

1. Bargmann-Montgomery-von Neumann 1946, Black-Southwell 1938, Cassinis 1946, Collatz 1942, 1948a, Hotelling 1943a, Householder 1950, Kantorovich 1948, Milne 1951, Seidel 1874, Taussky 1949, 1950, Temple 1939, von Mises-Pollaczek-Geiringer 1929, Zurmühl 1950.

2. Bodewig 1947b, Cesari 1937b.

3. Satterthwaite 1944.

c.

1.

2. Bodewig 1947b, Bowie 1950, Jürgens 1886.

d. Geiringer 1949, Hotelling 1936a, 1943a.

e. Jacobi 1845, Seidel 1874.

18. Akushsky 1946a, 1946b, 1946c, Alt 1946, Basile 1949, Eckert 1940, Engineering Research Associates 1950, Flanagan 1940, Forsythe 1951, Fox-Huskey-Wilkinson 1948b, Grossman 1948, Hartley 1946, IBM 1950 (complete), (IBM educational forums), Kormes 1943, Leppert 1947, Opler 1951, Petrie 1953, Renner 1946, Sherman 1951, Snedecor 1928, Tucker 1940, Verzuh 1949.

19. Engineering Research Associates 1950, Fröberg 1953, Goldstine-von Neumann 1951, Hartree 1949b, Harvard Computation Lab. 1946, McPherson 1948, Mitchell 1948, Rosser 1950, Snyder-Livingston 1949, Stiefel 1951, 1952.

19A. Adcock 1948, Anonymous 1934, Berry-Wilcox-Rock-Washburn 1946, Bowie (date unknown), Brown 1947, Dennis 1948, Engineering Research Associates 1950, Fairthorne 1944, Frame 1945, Fuchs 1914, Gorushkin 1948, Gradsteĭn 1947, Hardy-Dench 1948, Haupt 1950, Hazen-Schurig-Gardner 1930, Korn 1949, Kron 1945, Mallock 1933, Many 1950, McCann 1949, Meerovich 1947, Muirhead 1912, Murray 1947, 1949, Näbauer 1910, Parker 1941, 1945, Picchi 1948, Pipes 1948, Proshko 1947, Samssonow 1943, Schumann 1940, Spangenberg-Walters-Schott 1949, Wilbur 1936, zur Capellen 1948.

20.

a. Black (date unknown), Dedekind 1901, Cross 1932, Gauss 1823, Jacobi 1845, Salvadori 1948.

1. Forsythe 1951, Fox 1948, 1950a?, 1950b, Southwell 1940, 1946, Zurmühl 1950.

b. Aitken 1932, Albert 1941, Dwyer 1951, Fisher 1938, Forsythe 1951, Frazer-Duncan-Collar 1938, Hoel 1941, Hotelling 1943a, Householder 1950, Milne 1949, G. Schulz 1933.

c. Burgess 1916, Dwyer 1941c, 1951, Forsythe 1951.

d. Forsythe 1951.

e. Fox 1950a, Goodwin 1950, Pipes 1941, Zurmühl 1950.

21.

a.

1. Forsythe 1951, Fox 1948, Southwell 1940?, 1946?

b.

1. Aitken 1926, 1937b, 1950, Forsythe 1951, Fox 1950a, Samuelson 1945, Shanks 1949.

2. Forsythe 1951, Samuelson 1945, Shanks 1949.

22. Bargmann-Montgomery-von Neumann 1946, Bodewig 1947b, Cassina 1948, Collatz 1948a, Frankel 1950, Heck 1946, Kamela 1943, Parkes 1950, Ricci 1949.

23.

a. Bisshopp 1945, Good 1950.

b. Goldstine-von Neumann 1951, Jenne 1949, Kerkhofs 1947.

V. Ballantine 1931, W. E. Deming 1935, D. B. Duncan-Kenney 1946, Dunlap 1939, Dwyer 1941b, 1946, Encke 1835, Flanagan 1940, Fox 1947, 1950a, Fox-Huskey-Wilkinson 1948b, Griffin 1931, Grossman 1948, 1950, Halberstadt 1914, Hartley 1946, Heck 1946, Hendrikz (date unknown), Horst 1941a, Hruška 1943, Jenne 1949, Jordan 1920, Jossa 1938, 1940, Kantorovich-Krylov 1941, Kerkhofs 1947, Kolmogoroff 1946, Kormes 1943, 1945, Li͡usternik 1947, Magnier 1948, Oppokov 1939, Panov 1938, Picone 1940, Reich (date unknown), Schaefer 1927, H. Schulz 1938, Shaw 1946, Turton 1945, Wegner (date unknown).

V. Bibliography, With Cross-References to the Outline

In this section are given approximately 450 titles on the solution of linear equations, taken from the compiler's card file on numerical matrix methods. The card file has been collected during the past two years as part of an investigation at the National Bureau of Standards, Los Angeles, of old and new matrix methods suited to automatic digital computers.

The bibliography is most nearly complete in recent titles dealing with mathematical methods for obtaining A^{-1} or $A^{-1}b$, where A is a nonsingular square matrix. The bibliographies of Bodewig 1947b, Dwyer 1951, Harvard Computation Laboratory 1946, Higgins 1949, Hotelling 1943, Householder 1949, Kantorovich 1948, Kantorovich-Krylov 1948, and Ostrowski-Todd-Todd 1949 proved invaluable in starting and augmenting the list. Original papers were consulted when reasonably obtainable, and their references were added to the card file. Certain papers which appear to be covered in subsequent books have been omitted; some titles containing lists of such papers are given under heading 0c of section IV.

Because much of the research on solving linear equations has been ancillary to the solution of numerical problems in pure mathematics, statistics, physics, astronomy, geodesy, psychology, economics, engineering, etc., many good references are from these diverse fields. The present bibliography is less complete in these related fields, some of which have their own bibliographies (e. g., Higgins 1949 on numerical solution of partial differential equations).

There are several mathematical fields which are related to our subject, but which have not been included here. Three of these are known to be subjects for bibliographies now in preparation: Motzkin 1951 (on linear inequalities and applications); Taussky 1951 (on bounds for eigenvalues); Schwerdtfeger 1951 (on iteration in general; listed at end of bibliography).

At the last minute an effort was made to include some representative titles on analogue machinery for solving linear equations, starting with references in Frame 1945 and Tryon 1951. However, in regard to analogue machinery the bibliographies of Engineering Research Associates 1950 and zur Capellen 1948 are far more extensive than the few titles included below. Also, the bibliography of International Business Machines 1950 (mostly reproduced in Engineering Research Associates 1950) is more complete on uses of IBM equipment for matrix problems than the present bibliography.

The numbers in brackets after each title in the bibliography refer to the outline in section III, and summarize the contents of the reference. The notation 11b*1*(*B*&*F*&*I*) means that the paper treats the subcase of subject 11b*1*, in which *B*, *F*, and *I* are simultaneously fulfilled (logical product).

A circle (°) is placed before each title which the compiler has not examined personally, insofar as he can remember.

The abbreviations MR, Fs, Zb, SA after titles refer to abstract journals:

MR	=*Mathematical Reviews;*
Fs	=*Jahrbuch über die Fortschritte der Mathematik;*
Zb	=*Zentralblatt der Mathematik und ihre Grenzgebiete;*
SA	=*Science Abstracts, Section A.*

References to abstract journals are listed when they happen to be known; there is no uniform policy.

The following journal abbreviation has been used in the references: *MTAC, Mathematical Tables and Other Aids to Computation.*

Attention is invited to the supplementary bibliographies at the end, containing papers noted too late for inclusion in the main list.

Adams, Oscar S. 1915: *Application of the Theory of Least Squares to the Adjustment of Triangulation*, USCGS Special Publ. No. 28, U. S. Govt. Printing Office. [Iα?].

°Adcock, W. A. 1948: "An automatic simultaneous equation computer and its use in solving linear equations," *Rev. Sci. Inst.* **19,** 181–187. [19A].

Aitken, A. C. 1926: "On Bernoulli's numerical solution of algebraic equations," *Proc. Roy. Soc. Edinburgh* **46,** 289–305. [21b*1*].

Aitken, A. C. 1932: "On the evaluation of determinants, the formation of their adjugates, and the practical solution of simultaneous linear equations," *Proc. Edinburgh Math. Soc.* (2) **3,** 207–219. [2a, 2b, 20b].

Aitken, A. C. 1937a: "Studies in practical mathematics. I. The evaluation, with applications, of a certain triple product matrix," *Proc. Roy. Soc. Edinburgh* **57,** 172–181. [1a, 2c].

Aitken, A. C. 1937b: "Studies in practical mathematics. II. The evaluation of the latent roots and latent vectors of a matrix," *Proc. Roy. Soc. Edinburgh* **57,** 269–304. [21b*1*].

Aitken, A. C. 1945: "Studies in practical mathematics. IV. On linear approximation by least squares," *Proc. Roy. Soc. Edinburgh* **A62,** 138–146. [2a].

Aitken, A. C. 1950: "Studies in practical mathematics. V. On the iterative solution of a system of linear equations," *Proc. Roy. Soc. Edinburgh* **A63,** 52–60. [10a, 10c, 13c*1*, 21b*1*].

°Akushsky, I. J. 1946a: "Numerical solution of the Dirichlet equation with the aid of perforated card machines," *C. R. (Doklady) Acad. Sci. URSS* **52,** 375–378. [18].

°Akushsky, I. J. 1946b: "The four-counter scheme of solution of Dirichlet's problem by means of punched-card machines," *C. R. (Doklady) Acad. Sci. URSS* **54,** 659–662. [18].

°Akushsky, I. J. 1946c: "On numerical solution of Dirichlet problem on punched-card machines," *C. R. (Doklady) Acad. Sci. URSS* **54,** 755–758. [18].

Albert, A. A. 1941: "A rule for computing the inverse of a matrix," *Am. Math. Monthly* **48,** 198–199 (MR **2,** 243). [20b].

Alt, Franz L. 1946: "Multiplication of matrices," *MTAC* **2,** 12–13. [18].

°Andersen, Einar 1947: "Solution of great systems of normal equations together with an investigation of Andrae's dot-figure. An arithmetical-technical investigation," *Mem. Inst. Géodésique Danemark* (*Geodaetisk Inst. Skr.*) (3), **11,** 65 pp. (MR **9,** 622). [2a, 2c].

Andree, R. V. 1951: "Computation of the inverse of a matrix," *Am. Math. Monthly* **58,** 87–92. [2a, 2e, 20e].

°Anér, H. 1926: "Ausgleichung durch Anwendung des arithmetischen Mittels," *Zeitschrift für Vermessungswesen* **55,** 65–77. [10b].

°Anonymous? 1934: "The Mallock electrical calculating machine," Reprint from *Engineering* (London), June 22, 1934, 8 pp. [19A].

Argelander, F. W. A. 1844: "Ueber die Anwendung der Methode der kleinsten Quadrate auf einen besondern Fall," *Astr. Nachr.* **21,** No. 491, 163–168. [10. Method unclear.]

°Asplund, Lars 1945: "Über einige Methoden für die Ausgleichung geodätischer Netze," *Rikets Allmäna Kartverk*, Meddelande No. 5, Stockholm. [2c*1*].

Backman, Gaston 1946: "Rekursionsformeln zur Lösung der Normalgleichungen auf Grund der Krakovianenmethodik," *Arkiv för Matematik, Astronomi och Fysik* **33A,** 1–14. [2f].

Baetslé, P. L. 1951: "Systémisation des calculs numériques de matrices," *Bulletin Géodésique*, 22–41. [2a, 2c, 2e].

°Ballantine, J. P. 1931: "Numerical solutions of linear equations by vectors," *Amer. Math. Monthly* **38,** 275–277. [V].

°Banachiewicz, T. 1937a: "Zur Berechnung der Determinanten, wie auch der Inversen, und zur darauf basierten Auflösung der Systeme linearer Gleichungen," *Acta Astronomica*, **3,** 41–67. [4].

°Banachiewicz, T. 1937b: "Calcul des déterminants par la méthode des cracoviens," *Bull. Intern. de l'Acad. Polonaise, Série A, Sci. Math.*, 109–120. [2f, 4].

°Banachiewicz, T. 1937c: "Sur la résolution numérique d'un système d'équations linéaires," *Bull. Intern. de l'Acad. Polonaise, Série A, Sci. Math.*, 350–354. [2f, 4?].

°Banachiewicz, T. 1938a: "Principes d'une nouvelle technique de la méthode des moindres carrés," *Bull. Intern. de l'Acad. Polonaise, Série A, Sci. Math.*, 134–135. [2e*1*?].

°Banachiewicz, T. 1938b. "Méthode de résolution numérique des équations linéaires, du calcul des déterminants et des inverses, et de réduction des formes quadratiques," *Bull. Intern. de l'Acad. Polonaise, Série A, Sci. Math.*, 393–404. Reprinted in *Cracow Observatory Reprint* 22. [2c*1*].

Banachiewicz, T. 1942: "An outline of the Cracovian algorithm of the method of least squares," *Astron. Journal*, 38–41. [2f].

Banachiewicz, T. 1951: "Résolution d'un système d'équations linéaires algébriques par division," *Enseignement Math.* **39,** (1942–1950), 34–45. [2c, 2f].

Bargmann, V., Montgomery, D., and von Neumann, J. 1946: "Solution of linear systems of high order," Report prepared for the Bureau of Ordnance (Contract NORD–9596) (25 Oct. 1946), 86 pp. [0, 1, 2a, 3b, 4, 11c*1*, 11e*2*, 16a*2*, 16b*2*, 17b*1*, 22].

°Bartlett, M. S. 1951: "An inverse matrix adjustment arising in discriminant analysis," *Ann. Math. Stat.* **22,** 107–111. [16*a*2].

°Basile, R. 1949: "Résolution de systèmes d'équations linéaires algébriques et inversions de matrices au moyen des machines de mécanographie comptable. Complément pratique par R. Janin," *Office National d'Études et de Recherches Aéronautiques*, Paris, publ. no. 28, v+21 pp. (MR **11,** 692). [3a?, 18].

Bauschinger, Julius 1901: "Ausgleichungsrechnung," *Enc. d. Math. Wiss.* ID 2, 786–798. [2a, 10].

Benoit, Commandant 1924: "Note sur une méthode de résolution des équations normales etc.," International Geodetic and Geophysical Union, Association of Geodesy, *Bulletin Géodésique* (Toulouse), no. 2, 67–77. (Translation by Rainsford available in UMT file of *MTAC*. See *MTAC*, April 1951.) [2c*1*].

Berkson, Joseph 1936: "Significant figures in statistical constants," *Science* **84,** 437. [16a*2*].

Berry, Clifford E. 1945: "A criterion of convergence for the classical iterative method of solving linear simultaneous equations," *Ann. Math. Stat.* **16,** 398–400. [10c].

°Berry, C. E., Wilcox, D. E., Rock, S. M., and Washburn, H. W. 1946: "A computer for solving linear simultaneous equations," *J. Appl. Phys.* **17,** 262–272. [19A].

°Biezeno, C. B. 1924: "Zeichnerische Ermittlung der elastischen Linie eines federnd gestützten, statisch unbestimmten Balkens," *Zeitschr. f. Angew. Math. Mech.* **4,** 93–102. [10d].

°Bingham, M. D. 1941: "A new method for obtaining the inverse matrix," *J. Amer. Stat. Assoc.* **36,** 530–534. [6b*1*].

°Birman, M. Sh. 1950: "Einige Abschätzungen für die Methode des schnellsten Abstiegs" (Russian), *Uspekhi Matem. Nauk* **5,** no. 3 (37), 152–155. (Zb **38,** 80). [11a (*1* & *6*)].

Bisshopp, K. E. 1945: "The inverse of a stiffness matrix," *Quart. Appl. Math.* **3,** 82–84. [1b, 23a].

Black, A. N. 1938: "The method of relaxation applied to survey problems," *Empire Survey Review* **4** (no. 29), 406–413. [11b1(*A* & *F* & *I*)?].

Black, A. N. (date unknown): "Approximate methods of solving normal equations," *Empire Survey Review* **7** (no. 52), 242–245. (Probably c. 1942.) [10b, 11b*1*(*A* & *F* & *I*), 11b*2*, 20a].

Black, A. N. 1949: "Further notes on the solution of algebraic linear simultaneous equations," *Quart. J. Mech. Appl. Math.* (Oxford) **2,** 321–324 (MR **11,** 743). [2a*1A*].

Black, A. N., and Southwell, R. V. 1938: "Relaxation methods applied to engineering problems. II. Basic theory, with application to surveying and to electrical networks, and an extension to gyrostatic systems," *Proc. Roy. Soc. A* **164,** 447–467. [11b*1*, 17b*1*].

°Blumenthal, O. 1914: "Über die Genauigkeit der Wurzeln linearer Gleichungen," *Zeit. Math. u. Physik* **62,** 359–362 (Fs **45,** 174). [16a*2*].

°Bodewig, E. 1947a: "Comparison of some direct methods for computing determinants and inverse matrices," *Nederl. Akad. Wetensch., Proc.* **50,** 49–57 (MR **8,** 407). [1a, 4].

Bodewig, E. 1947b: "Bericht über die verschiedenen Methoden zur Lösung eines Systems linearer Gleichungen mit reellen Koeffizienten. I, II, III, IV, V," *Nederl. Akad. Wetensch., Proc.* **50,** 930–941, 1104–1116, 1285–1295 and **51,** 53–64, 211–219. Same articles in *Indagationes Math.* **9,** 441–452, 518–530, 611–621 (1947), and **10,** 24–35, 82–90. [0, 2a, 2b, 2c, 3a*1*, 3b, 4, 5, 10a, 10b, 10c, 10d, 10f, 11c*1*, 11c*3*, 13a (*2* & *6*), 17b*2*, 17c*2*, 22].

Bodewig, E., and Zurmühl, R. 1950: "Zu R. Zurmühl: Zur numerischen Auflösung linearer Gleichungssysteme nach dem Matrizenverfahren von Banachiewicz. Z. angew. Math. Mech. **29** (1949) 76–84," *Zeitschr. f. Angew. Math. Mech.* **30,** 130–132. [2c].

°Boltz, H. 1923: "Entwicklungsverfahren zur Ausgleichung geodätischer Netze nach der Methode der kleinsten Quadrate," *Veröffentlichungen des Preussischen Geodätischen Instituts N. F.* no. 90, Berlin. [4].

Booth, A. D. 1949: "An application of the method of steepest descents to the solution of systems of nonlinear simultaneous equations," *Quart. J. Mech. Appl. Math* **2,** 460–468 (MR **11,** 693). [11a (*2* & *5*)].

Boschan, Paul 1946: "The consolidated Doolittle technique," (abstract) *Ann. Math. Stat.* **17,** 503. [2b].

Bottema, O. 1950: "A geometrical interpretation of the relaxation method," *Quart. Appl. Math.* **7,** 422–423. [13a (2 & *6*)].

°Bowie, O. L. (date unknown): "Electrical computing board for the numerical solution of partial differential equations," Watertown Arsenal Laboratory Report WAL 790/22. [19A].

Bowie, O. 1947: "Least-square application to relaxation methods," *J. Appl. Phys.* **18,** 830–837. [11b*1*?].

Bowie, O. L. 1950: "Practical solution of simultaneous linear equations," 13 Feb. 1950 report on O. O. Project No. TR 3-3027B, Watertown Arsenal, Mass. [17c*2*, 10c].

Brand, Louis 1935: "The method of moment distribution for the analysis of continuous structures," *Bull. Amer. Math. Soc.* **41,** 901–906. [10b?].

Brown, G. W. 1947: "The stability of feedback solutions of simultaneous linear equations," *Bull. Amer. Math. Soc.* **53,** 61 (abstract). [19A].

Bruner, Nancy 1947: "Note on the Doolittle solution," *Econometrica* **15,** 43–44. [2a*1A*].

Bückner. Hans 1950: "Über ein unbeschränkt anwendbares Iterationsverfahren für Systeme linearer Gleichungen," *Arch. Math.* **2,** 172–177. (MR **11,** 743). [13b].

Burgess, H. T. 1916: "On the matrix equation $BX=C$," *Am. Math. Monthly* **23,** 152–155. [20c].

°Cassina, Ugo 1948: "Sul numero delle operazioni elementari necessarie per la risoluzione dei sistemi di equazioni lineari," *Boll. Un. Mat. Ital.* (3) **3,** 142–147 (MR **10,** 405). [1a, 2a, 2c, 22].

°Cassinis, Gino 1944: "I metodi di H. Boltz per risoluzione dei sistemi di equazioni lineari e il loro impeigo nella compenzazioni della triangolazione," *Revista Catasto e Servizi Tecnici Erariali* No. 1. [4?].

°Cassinis, G. 1946: "Risoluzione dei sistemi di equazioni algebriche lineari," *Rend. Sem. Math. Fis. Milano* **17,** 62–78 (MR **9,** 622). [2a, 2c, 3b, 10b, 10c, 17b*1*].

Cauchy, A. L. 1847: "Méthode générale pour la résolution des systèmes d'équations simultanées," *Comptes Rendus Acad. Sci. Paris* **25,** 536–538. [11a (*2* & *5*), 11a (*2* & *4* & *5*)].

°Cesari, L. 1937a: "Sulla risoluzione dei sistemi di equazioni lineari per approssimazioni successive," *Rendic. Reale Accademia Nazionale dei Lincei, Classe Scienze Fis., Mat., Natur.* **25,** ser. 6a, Roma, 422–428. [10a?].

°Cesari, Lamberto 1937b: "Sulla risoluzioni dei sistemi di equazioni lineari per approssimazioni successive," Extract of *Rass. Poste, Teleg. e Telef.* **4,** 37 pp. (Zb **17,** 367). [10a, 17b*2*].

Chen, Pei-Ping 1944: "Dyadic analysis of space rigid frame-work," *J. Franklin Inst.* **238,** 325–334. [10b?].

°Cherepkov, F. S. 1936: "On the solution of systems of linear equations by the method of iteration," *Matem. Sbornik N. S.* **1** (43), 953–960 (Russian, Fr. extract) (Fs **62,** 1391). [10a?].

°Chiò, F. 1853: *Mémoire sur les Fonctions Connues sous le Nom de Résultants ou de Déterminants*, Turin. (Cf. Whittaker and Robinson 1924, p. 71.) [1a].

°Cimmino, Gianfranco 1938: "Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari," *Ricerca Scientifica*, Roma (2), **9**I, 326–333. Also in *Pubblicazioni dell'Istituto per le Applicazioni del Calcolo* no. 34 (Fs **64,** 1241) [10f].

°Clasen, B. I. 1888: "Sur une nouvelle méthode de résolution des équations linéaires et sur l'application de cette méthode au calcul des déterminants," *Ann. de la Société Scientifique de Bruxelles* **12,** A 50–59, B 251–281. [3a*2*].

°Cochran, W. G. 1938: "The omission or addition of an independent variate in multiple linear regression," *Suppl., J. Roy. Stat. Soc.* **5,** 171–176. [2b?].

Collar, A. R. 1939: "On the reciprocation of certain matrices," *Proc. Roy. Soc. Edinburgh* **59,** 195–206. (Inverting persymmetric matrices.) [1b].

Collar, A. R. (*see also* Frazer, R. A.)

Collatz, L. 1942: "Fehlerabschätzung für das Iterationsverfahren zur Auflösung linearer Gleichungssysteme," *Zeitschr. f. Angew. Math. Mech.* **22,** 357–361 (MR **5,** 50). [10b, 10c, 17b*1*].

Collatz, L. 1948a: "Graphische und numerische Verfahren," *FIAT Review of German Science, 1939–1946*. Applied Mathematics, Part I, O. M. G., Wiesbaden, 1–92 by Collatz. Especially 21–26. Bibliography. [1a*1*, 2a, 2b, 4, 10b, 10c, 17b*1*, 22].

Collatz, Lothar 1949: "Eigenwertaufgaben mit technischen Anwendungen," Leipzig (vol. 19, series A in *Math. und ihre Anwendungen in Physik und Technik*, Kamke and Kratzer, editors). 322–324. (16a*2*, 10b, 10c].

Collatz, L. 1950a: "Zur Herleitung von Konvergenzkriterien für Iterationsverfahren bei linearen Gleichungssystemen," *Zeitschr. f. Angew. Math. Mech.* **30,** 278–280 (Zb **37,** 359). [10b, 10c].

Collatz, L. 1950b: "Über die Konvergenzkriterien bei Iterationsverfahren für lineare Gleichungssysteme," *Math. Zeitschr.* **53,** 149–161 (Zb **38,** 77). (10a, 10b, 10c, 16b*1*].

Couffignal, Louis 1944: "Recherches de mathématiques utilisables. La résolution numérique des systèmes d'équations linéaires. I. L'opération fondamentale de réduction d'un tableau," *Revue Sci.* (*Rev. Rose Illus.*) **82,** 67–78 (MR **8,** 128). [2a?].

Cross, Hardy 1932: "Analysis of continuous frames by distributing fixed-end moments," Paper No. 1793, *Trans. Amer. Soc. Civil Engrs.* **96,** 1–10. (Reprinted in Grinter 1949.) [10b, 20a].

Crout, Prescott D. 1941: "A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients," *Trans. Amer. Inst. Elec. Engrs.* **60,** 1235–1240. (Reprinted by Marchant Calculating Machine Co. as Report MM–182.) [2c].

Curry, Haskell B. 1944: "The method of steepest descent for non-linear minimization problems," *Quart. Appl. Math.* **2,** 258–261. [11a (*2* & *5*)].

Dantzig, George B. 1951: *A Preliminary Note on Solving Linear Equations by the Revised Simplex Procedure*, Dittoed by Hq USAF, DCS/Comptroller, 20 July 1951, 5 pp. [7b*1*].

Dedekind, R. 1901: "Gauss in seiner Vorlesung über die Methode der kleinsten Quadrate," Festschrift zur Feier des 150-jährigen Bestehen der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Berlin, 45–59. Also in *Gesammelte Math. Werke* **2** (1931), 293–306. [11b*1*(*B* & *F* & *I*), 20a].

de la Garza, A. 1951: *An Iterative Method for Solving Systems of Linear Equations*, Carbide and Carbon Chem. Div. Union Carbide, K–25 Plant Report K–731, 15 Feb. 1951. [13a (*1* & *4*)].

°Deming, H. G. 1928: "A systematic method for the solution of simultaneous linear equations," *Amer. Math. Monthly* **35,** 360–363. [2a].

°Deming, W. Edwards 1937: (title unknown), *Science*, **85,** 451–454. [16a*2*?].

°Deming, W. Edwards 1938: *Some Notes on Least Squares*, USDA Graduate School, pp. 105, 111, 121, 135. [V].

Dench, E. C. (*see* Hardy, A. C.)

Dennis, Paul 1948: "Description of the 4-element machine," Mimeographed notes, Engineering extension, UCLA. [19A].

Dickson, L. E. 1922: *First Course in the Theory of Equations*, Wiley, p. 134. [1a].

°Dodgson, C. L. 1866: "Condensation of determinants," *Proc. Roy. Soc.* **15,** 150–155. [2a*2*].

°Doolittle, M. H. 1878: "Method employed in the solution of normal equations and the adjustment of a triangulation," *USCGS Report*, 115–120. [2a*1A*].

Dreyer, H. J. 1943: "Solution of systems of linear equations by means of punched-card machines," (translated by H. R. Grumman), A. M. C. translation F–TS–1046–RE, (MR **11,** 693). [3a*2*].

°Duncan, D. B., and Kenney, J. F. 1946: *On the Solution of Normal Equations and Related Topics*, Edwards Bros., 35 pp. [2c*1*, V].

Duncan, W. J. 1944: "Some devices for the solution of large sets of simultaneous linear equations (with an appendix on the reciprocation of partitioned matrices)," *Philos. Mag.* (7), **35,** 660–670 (MR **7,** 84). [4].

Duncan, W. J. (*see also* Frazer, R. A.)

Dunlap, J. W. 1939: *Workbook in Statistical Method*, Prentice-Hall, N. Y. [0c, V].

Dwyer, Paul S. 1941a: "The solution of simultaneous equations," *Psychometrika* **6,** 101–129. [0c, 2a].

Dwyer, P. S. 1941c: "The evaluation of linear forms," *Psychometrika* **6,** 355–365. [20c].

Dwyer, Paul S. 1941d: "The Doolittle technique," *Ann. Math. Stat.* **12,** 449–458. [2a*1A*].

Dwyer, Paul S. 1942: "Recent developments in correlation technique," *J. Amer. Stat. Assoc.* **37,** 441–460. [0c, 2c*1*].

Dwyer, P. S. 1944: "A matrix presentation of least squares and correlation theory with matrix justification of improved methods of solution," *Ann. Math. Stat.* **15,** 82–89 (MR **5,** 245). [2c*1*].

Dwyer, Paul S. 1945: "The square root method and its use in correlation and regression," *J. Amer. Stat. Assoc.* **40,** 493–503 (MR **7,** 338). [2c*1*].

°Dwyer, Paul S. 1946: "Simultaneous computation of correlation coefficients with missing variates," *Proc. of the Research Forum,* 20–27, N. Y., IBM, Aug. 1946. [V].

Dwyer, Paul S. 1951: *Linear Computations,* Wiley, 344 pp. [0b, 0c, 1a*1*, 2a, 2b, 2c, 2d, 2e, 6a, 6b*1*, 16a*2*, 16b*2*, 16b*3*, 20b, 20c].

Dwyer, P. S. 1953: "Errors of matrix computations," *Simultaneous Linear Equations and the Determination of Eigenvalues.* NBS AMS29 (1953). [16].

Dwyer, P. S. (*see also* Waugh, F. V.)

Eckert, W. J. 1940: *Punched Card Methods in Scientific Computation,* Columbia Univ. [18].

°Encke, H. 1835: (title unknown), *Astronomisches Jahrbuch* (1835), 267–272 and (1836), 263. [2a, V].

Engineering Research Associates 1950: *High-Speed Computing Devices,* McGraw-Hill. [0c (machines), 18, 19, 19A].

Etherington, I. M. H. 1932: "On errors in determinants," *Proc. Edinburgh Math. Soc.* (2) **3**, 107–117. [16a*2*].

Ezekiel, Mordecai (see Tolley, H. R.).

°Fairthorne, R. A. 1944: "Mechanical instruments for solving linear simultaneous equations," *Aeronaut. Res. Council R. and M. 2144.* [19A].

Fisher, R. A. 1938: *Statistical Methods for Research Workers,* London, 7th ed., p. 256. [20b].

°Flanigan, John C. 1940: "A successive approximation solution for prediction problems involving a large number of variables," *Proc. of the Educational Research Forum,* 75–79, IBM, N. Y. [18, V].

Flanders, Donald A., and Shortley, George 1950: "Numerical determination of fundamental modes," *J. Appl. Phys.* **21**, 1326–1332. [15a, applied to eigenvalue problems].

Forsythe, George E. 1951: "Theory of selected methods of finite matrix inversion and decomposition," Lectures in Math. 136, notes by D. G. Aronson and K. Iverson. INA Report 52–5, 93 pp. [0, 2a. 2c, 2e, 3a*2*, 3a*3*, 3b, 4, 6b, 10a, 10b, 10c, 10d, 10f, 11a (comprehensive), 11b*1* (*A* & *F* & *I*), 12a, 12e, 16a*2*, 18, 20a*1*, 20b, 20c, 20d, 21a, 21b].

Forsythe, George E., and Leibler, Richard A. 1950: "Matrix inversion by a Monte Carlo method,"*MTAC* **4**, 127–129 and **5** (1951), 55. [15c*1*].

Forsythe, G. E., and Motzkin, T. S. 1950: "On a gradient method for solving linear equations," multilithed outline at INA. [11a (*2* & *5*), 11a (3 & *5*)].

Forsythe, G. E., and Motzkin, T. S. 1951a: "Asymptotic properties of the optimum gradient method," abstract, *Bull. Amer. Math. Soc.* **57**, 183. [11a (*2* & *5*)].

Forsythe, George E., and Motzkin, Theodore S. 1951b: "Acceleration of the optimum gradient method. Prelim. report," abstract *Bull. Amer. Math. Soc.* **57,** 304–305. [11a (*2* & *5*), 11a (*3* & *5*), 12c].

Forsythe, George E., and Motzkin, Theodore S. 1952: "An extension of Gauss' transformation for improving the condition of systems of linear equations," *MTAC* **6,** 9–17. [17a*1*].

Fox, L. 1947: "Some improvements in the use of relaxation methods for the solution of ordinary and partial differential equations," *Proc. Roy. Soc. A* **190**, 31–59. [11b*1*, V].

Fox, L. 1948: "A short account of relaxation methods," *Quart. J. Mech. Appl. Math.* **1**, 253–280 (MR **10**, 574). [11b*1* (*A* & *F* & *I*), 11b*1* (*B* & *F* & *I*), 11b*1L*, 11b*2*, 20a*1*, 21a].

Fox, L. 1950a: "Linear equations and reciprocal matrices," N. P. L., Math. Div., Methodology Progress Report No. 1, photostat in INA library, 59 pp. (Submitted to *J. of Research.*) [0, 2a, 2a*1*, 2b*1*, 2c, 2d, 2e, 3a*2*, 3b, 4, 10b, 10c, 11b*1* (*C* & *F* & *I*), 11c*1*, 20e, 21b*1*].

Fox, L. 1950b: "Practical methods for the solution of linear equations and the inversion of matrices," *J. Roy. Stat. Soc.*, Ser. **12**, B, 120–136 (Zb **38, 77**). [2c, 11b*1*, (*A* & *G* & *J*), 11b*1* (*E* & *I* & *L*), 12f, 20a*1*].

Fox, L., Huskey, H. D., and Wilkinson, J. H. 1948a: "Notes on the solution of algebraic linear simultaneous equations," *Quart. J. Mech. Appl. Math.* **1**, 149–173. Translated into Russian, *Uspekhi Matem. Nauk* **5** (1950), No. 3, 60–86 (MR **11**, 743). [2a, 2d, 2b*1*, 2c, 3a*1*].

Fox, L., Huskey, H. D., and Wilkinson, J. H. 1948b?: "The solution of algebraic linear simultaneous equations by punched card methods," mimeographed at N. P. L. Supplements Fox-Huskey-Wilkinson 1948a. [18, V].

Frame, J. S. 1945: "Machines for solving algebraic equations," *MTAC* **1**, 337–353. [0c, 19A].

Frame, J. S. 1949: "A simple recursion formula for inverting a matrix," abstract *Bull. Amer. Math. Soc.* **55**, 1045. [6b*1*].

Frankel, Stanley P. 1950: "Convergence rates of iterative treatments of partial differential equations," *MTAC* **4**, 65–75. [10d, 10g, 13a (*1* & *5*) ≡ 10c, 13a (*1* & *3* & *5*), 22].

Frazer, R. A 1947: "Note on the Morris escalator process for the solution of linear simultaneous equations," *Philos. Mag.* (7) **38**, 287–289. [2b].

Frazer, R. A., Duncan, W. J., and Collar, A, R. 1938: *Elementary Matrices and Some Applications to Dynamics and Differential Equations,* Cambridge Univ. Press, chap. 4. [0, 2a, 2d, 3a*1*, 4, 10a, 10b, 10c, 11c*1*, 20b].

Freeman, G. F. 1943: "On the iterative solution of linear simultaneous equations," *Philos. Mag.* (7) **34**, 409–416. [6d, 6b*2*].

Fried, Bernard (see Shortley, George H.).

Fröberg, C. E. 1953: "Solutions of linear systems of equations on a relay machine," *Simultaneous Linear Equations and Determination of Eigenvalues.* NBS AMS29. [2a, 19].

°Fuchs, K. 1914: "Hydrostatische Gleichgewichtsmaschinen," *Zeitschr. Math. Phys.* **63**, 203–214. [19A].

Gardner, M. F. (see Hazen, H. L.).

Gaskell, R. E. 1943: "On moment balancing in structural dynamics," *Quart. Appl. Math.* **1**, 237–249. [11b*1* (*A* & *F* & *I*)?].

Gatto, Franco 1949: "Sulla risoluzione numerica dei sistemi di equazioni lineari," *Ricerca Scientifica* **19**, 1385–1388. [10c].

Gauss, C. F. 1823: "Letter to Gerling, 26 Dec. 1823," *Werke* **9**, 278–281. Translated by G. E. Forsythe, *MTAC* **5**, 255–258, under title "Gauss to Gerling on Relaxation". Reprinted in Schaefer 1927. [11b*1* (*B* & *F* & *I*), 16b*3*, 20a].

Gauss, C. F. 1826: "Supplementum theoriae combinationis observationum erroribus minimis obnoxiae," *Werke*, Göttingen, **4**, 55–93. [2a, 10c*1*].

Geiringer, Hilda P. 1942: "On the numerical solution of linear problems by group iteration," *Bull. Amer. Math. Soc.* **48**, 370. [10e].

Geiringer, Hilda 1949: "On the solution of systems of linear equations by certain iteration methods," *Reissner Anniversary Volume, Contrib. to Appl. Mech.* 365-393. Edwards Bros. [0a, 10a, 10b, 10c, 10d, 10e*1*, 13c, 14b, 17d].

Geiringer, Hilda: (See also Pollaczek-Geiringer, Hildă and von Mises, R.).

Gerling, Christian Ludwig 1843: *Die Ausgleichungs-Rechnung der practischen Geometrie*, Hamburg and Gotha, (U. of Ill. library). [11b*1* (*B* & *F* & *I*), 11b*1L*, 17a*1*].

Goldstein, L. (see also Nielsen, K. L.).

Goldstine, Herman H., and von Neumann, John 1951: "Numerical inverting of matrices of high order, II," *Proc. Amer. Math. Soc.* **2**, 188–202. (Part I under von Neumann and Goldstine.) [2a*1*, 16b*2*, 19, 23b].

Goodwin, E. T. 1950: "Note on evaluation of complex determinants," *Proc. Cambr. Phil. Soc.* **46**, 450–452. [1a*1*, 20c].

°Gorushkin, V. I. 1948: "Linear transformation of coordinates in the theory of electric machines and matrix calculus," *Izv. Akad. Nauk SSSR, Otd. Tekhn. Nauk*, 533–544. [19A?].

°Gradshteĭn, I. S. 1947: "The solution of systems of linear equations by L. I. Gutenmakher's electrical models," *Izv. Akad. Nauk SSSR, Otd. Tekhn. Nauk*, 529–584 (MR **9**, 210). [19A].

Gram, J. P. 1883: "Ueber die Entwickelung reeller Functionen in Reihen mittelst der Methode der kleinsten Quadrate," *J. Reine Angew. Math.* **94**, 41–73. [4].

°Griffin, H. D. 1931: "On partial correlation versus partial regression for obtaining the multiple regression equations," *J. Educ. Psych.* **22**, 35-44. [0c, V].

Grinter, L. E. (editor) 1949: *Numerical Methods of Analysis in Engineering*, MacMillan. (Contains Higgins 1949.) [0c].

°Grossman, D. P. 1948: "The application of punched-card machines to the solution of a system of linear algebraic equations by the iteration method," *Izv. Akad. Nauk SSSR, Otd. Tekhn. Nauk*, 1229-1238 (Russian) (MR **10**, 574). [10, 18, V].

°Grossman, D. P. 1950: "On the problem of the numerical solution of systems of compatible linear algebraic equations," *Uspekhi Matem. Nauk* **5**, No. 3, 87–103. [0?, V].

Guttman, Louis 1946: "Enlargement methods for computing the inverse matrix," *Ann. Math. Stat.* **17**, 336–343. [4].

°Halberstadt, S. 1914: "Zur Methode der kleinsten Quadrate," *Gött. Nachr. Math. Phys. Kl.*, 309-323. [V].

°Hardy, Arthur C., and Dench, Edward C. 1948: "An electronic method for solving simultaneous equations," *J. Opt. Soc. Amer.* **38**, 308–312. [19A].

Hartley, H. O. 1946: "The application of some commercial calculating machines to certain statistical calculations," *Suppl. J. Roy. Stat. Soc.* **8**, 154-173; discussion 173-183 (MR **9**, 251). [18, V].

Hartree, D. R. 1948: "Experimental arithmetic," *Eureka* **10**, 13-18. [11a (*2* & *5*), 11a*4*?, 16a*1*].

Hartree, Douglas R. 1949b: *Calculating Instruments and Machines*, Urbana, pp. 119ff. [0, 11a (*2* & *5*), 19].

Harvard Computation Laboratory 1946: *A Manual of Operation for the Automatic Sequence Controlled Calculator*, Harvard. Univ. Press. [2a, 19].

°Haupt, L. M. 1950: "Solution of simultaneous equations through use of the a. c. network calculator," *Rev. Sci. Inst.* **21**, 683-686 (SA 10 (1951)). [19A].

°Hazen, H. L., Schurig, O. R., and Gardner, M. F. 1930: "The M. I. T. network analyzer," *Trans. Amer. Inst. Elec. Engrs.* **49**, 1102- [19A].

°Heck, O. 1946: "Über den Zeitaufwand für das Berechnen von Determinanten und für das Auflösen von linearen Gleichungen," *Diss. Tech. Hochschule Darmstadt.* [22, V].

Hellinger, Ernst, and Toeplitz, Otto 1924: "Integralgleichungen und Gleichungen mit unendlichvielen Unbekannten," II C 13 of *Encyklopädie der Math. Wiss.* [10d].

Hendrikz, Deryck R. (date unknown): "Relaxation and the coordinate method of triangulation adjustment," *Empire Survey Review* **5** (no. 36), 358 363. (Probably c. 1940.) [11b*1*?, V].

°Hertwig, A. 1905: "Beziehungen zwischen Symmetrie und Determinanten in einigen Aufgaben der Fachwerktheorie," *Festschrift Adolph Wüllner*, 194 213. [16a?].

°Hertwig, A. 1912: "Die Lösung linearer Gleichungen durch unendliche Reihen und ihre Anwendungen auf die Berechnung hochgradig statisch unbestimmter Systeme," *Festschrift für H. Müller-Breslau*, Leipzig, 37-59. [10c*1*].

Herzberger, M. 1949: "The normal equations of the method of least squares and their solution," *Quart. Appl. Math.* **7**, 217-223. (Vector interpretation of elimination method.) [2a, 2c, 2d].

Herzberger, M., and Morris, R. H. 1947: "A contribution to the method of least squares," *Quart. Appl. Math.* **5**, 354-357. [2a, 2d].

Hestenes, M. R. 1951: "Iterative methods for solving linear equations," NAML Report 52-9. [6d].

Higgins, Thomas J. 1949: "A survey of the approximate solution of two-dimensional physical problems by variational methods and finite difference procedures," chap. 10 of Grinter 1949. (Essentially annotated bibliography of 140 titles. Mostly partial differential equations. Many Russian titles.) [10c].

Hoel, Paul G. 1940: "The errors involved in evaluating correlation determinants," *Ann. Math. Stat.* **11,** 58–65. [16b*2A*, 16b*2B*].

Hoel, Paul G. 1941: "On methods of solving normal equations," *Ann. Math. Stat.* **12,** 354–359. [2a, 2c, 20b].

Holley, Julian L. 1951: "Note on the inversion of the Leontief matrix," *Econometrica* **19,** 317-320. [10d].

°Horst, Paul 1941a: "A note on a machine method for the quantification of attributes," *The Prediction of Personal Adjustment*, Bull. 48 of Social Science Research Council, 347-348, N. Y. [V].

Hotelling, Harold 1933: "Analysis of a complex of statistical variables into principal components," *J. Educ. Psych.* **24,** 417-441, 498-520. [10c].

Hotelling, Harold 1936a: "Simplified calculation of principal components," *Psychometrika* **1,** 27-35. [17d].

Hotelling, Harold 1943a: "Some new methods in matrix calculation," *Ann. Math. Stat.* **14,** 1-34 (expository, bibliography). [0, 4, 6a*1*, 10a, 10c, 10d, 11c*1*, 16a*2*, 16b*2A*, 16b*2B*, 17b*1*, 17d, 20b].

Hotelling, Harold 1943b: "Further points on matrix calculation and simultaneous equations," *Ann. Math. Stat.* **14,** 440-441. [Supplements 1943a.]

Hotelling, H. 1949: "Practical problems of matrix calculation," *Proc. Berkeley Symposium on Math. Stat. and Prob.*, 1945–6, 275–293. [0, 4, 11c*1*, 16b].

Householder, A. S. 1949: "Notes on numerical methods," Multilithed typescript, Oak Ridge. [0, 2a, 2b, 2c, 3a*2*, 10a, 10b, 10c, 10d, 11a(comprehensive), 11b*1*(*C* & *F* & *I*), 11b*1*(*D* & *G* & *J*), 11c*1*, 13a(comprehensive), 16b].

Householder, A. S. 1950: "Some numerical methods for solving systems of linear equations," *Amer. Math. Monthly* **57,** 453-459. [0, 2a, 2b, 2c, 3b, 10b, 10c, 10f, 11a(*1* & *6*), 11b*1*(*A* & *F* & *I*), 13a(*1* & *5*), 13a(*1* & *4*), 13a(*2* & *4*), 17b*1*, 20b].

°Hruška, Václav 1943: "Lösung von Gleichungssystemen durch das Iterationsverfahren," *Acad. Tchèque. Sci. Bull. Cl. Sci. Math. Nat.* **44,** 239–304, 399–422. [V].

Huskey, H. D. (see Fox, L.)

Ingraham, M. H. 1937: "A note on determinants," *Bull. Amer. Math. Soc.* **43,** 579-580. [1a, 4].

International Business Machines Corp. 1950: *Bibliography on the Use of IBM Machines in Science, Statistics, and Education*, IBM Corp., N. Y. [0c, 18].

Istituto per le Applicazioni del Calcolo, Consiglio Nazionale delle Ricerche (date unknown): "Rizoluzione di un particolare sistema di equazioni algebriche lineari," pub. no. 3, 3 pp. [2c*1*].

Istituto per le Applicazicne del Calcolo, Consiglio Nazionale delle Ricerche (date unknown): Inversione di una matrice quadrata simmetrica di ordine 24," pub. no. 4, Roma. 3 pp. [2c].

°Ivanov, V. 1939: "On the convergence of the process of iteration in the solution of a system of linear algebraic equations," (Russian, English summary.) *Bull. Acad. Sci. URSS*, Ser. Math. (*Izv. Akad. Nauk SSSR*), 477–483 (MR **2,** 118). [10b 10c].

Jacobi C. G. J. 1845: "Ueber eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommenden lineären Gleichungen," *Astr. Nachr.* **22,** no. 523, 297–306. Jacobi's *Werke* **3,** 467? [10b, 17e, 20a].

°Janet, M. 1920a: "Sur les systèmes d'équations aux dérivées partielles," *Comptes Rendus Acad. Sci. Paris,* **170,** 1101–1103, and *J. de Math.* (8) **3,** 65–151 (Fs **47,** 440). [16a*2*].

°Janet, M. 1920b: "Sur les systèmes d'équations aux dérivées partielles et les systèmes de formes algébriques," *Comptes Rendus Acad. Sci. Paris* **170,** 1236–1239 (Fs **47,** 440). [16a*2*?].

Jenne, W. 1949: "Zur Auflösung linearer Gleichungssysteme," *Astr. Nachr.* **278,** 73–95 (MR **11,** 692). Solves normal equations after K. Friedrich (geodetist). Tool: inversion of Jacobi matrices by continued fractions. Related polynomials of Nörlund studied. (0c, 1b, 23b, V].

Jensen, Henry 1944: "An attempt at a systematic classification of some methods for the solution of normal equations," *Geodaetisk Institut, Meddelelse* No. 18, Copenhagen. [0b, 1a*1*, 2a, 2c, 3a*2*, 4].

Jensen, Henry 1948: "On the superposition of the differential-equations of the geodetic line. With a calculation-example," *Geodaetisk Inst. Skr.* (3) **13,** 23 pp. [2a].

°Jordan, W. 1920: *Handbuch der Vermessungskunde*, vol. I, 7th edit., Stuttgart. p. 36. [3a*2*, V].

°Jossa, F. 1938: "Metodo di influenza per il calcolo di strutture iperstatiche mediante la successiva introduzione dei vincoli," *Ricerche Ingegneria* **6,** 61–65 (Fs **64,** 1434). [2b, 4, V].

°Jossa, Franco 1940: "Risoluzione progressiva di un sistema di equazioni lineari," *Rend. Accad. Sci. Fis. Mat. Napoli* (4) **10,** 346–352 (Fs **66,** 576) (MR **8,** 535). [2b, 4, V].

Jürgens, Enno 1886: *Zur Auflösung linearer Gleichungssysteme und numerischen Berechnung von Determinanten,* Festgabe. Aachen. Palm. (Fs **18,** 62). [2a, 3a*1*, 11b*1*(*B* & *F* & *I*), 16a*1*, 17c*2*].

°Kaczmarz, S. 1937: "Angenäherte Auflösung von Systemen linearer Gleichungen," *Bull. Intern. de l'Acad. Polonaise, Série* A, 355–357. [12a (*2* & *6*)].

°Kamela, Czesław 1943: "Die Lösung der Normalgleichungen nach der Methode von Prof. Dr. T. Banachiewicz," *Schw. Zeitschr. f. Verm. und Kulturtech.* **41,** 225–232 and 265–275 (MR **7,** 488). [2f, 22].

Kantorovitch, L. 1939: "The method of successive approximations for functional equations," *Acta Math.* **71,** 62–97 (English) (MR **1,** 18). [10].

Kantorovitch, L. V. 1945: "On an effective method of solving an extremal problem for quadratic functionals," *Comptes Rendus (Doklady) Acad. Sci. URSS (N.S.)*, **48,** 455–460 (English). [11a *(1 & 6)*].

Kantorovich, L. V. 1947: "On the method of steepest descent," *Doklady Akad. Nauk* **56,** 233–236 (MR **9**, 308) (Russian). [11a *(1 & 6)*, 11a *(3 & 6)*].

Kantorovich, L. V. 1948: "Functional analysis and applied mathematics," *Uspekhi Matem. Nauk N. S.* **3,** No. 6, 89–185 (MR **10,** 380). Also *Vestnik Leningrad Univ.* **3,** No. 6, 3–18 (1948)? (Both Russian.) [11a *(2 & 5)*, 17b*1*]. A translation is being issued by the NBS, Los Angeles.

°Kantorovich, L. V. 1949: "On Newton's method," *Trudy Mat. Inst. Steklov* **28,**, 104–144 (MR **12**, 419). (Russian) [11c*1*].

°Kantorovich, L. V., and Krylov, V. I. 1941: *Approximate Methods of Higher Analysis*, Leningrad-Moscow, (GTTI), 618 pp. A third edition (1950) has 695 pp. (Russian) [V].

Kantorovich, L. V., and Krylov, V. I. 1948: "Approximation Methods", in section on numerical methods of *Math. in USSR in the Thirty Years 1917–1947*, Moscow-Leningrad, 759–801 (Russian) (Zb **38,** 75). [0c, 10c, 11a *(2 & 5)*].

Kelley, Truman L., and Salisbury, Frank S. 1926: "An iteration method for determining multiple correlation constants," *J. Amer. Stat. Assoc.* **21,** 282–292. [13c].

Kenney, J. F. (see Duncan, D. B.)

°Kerkhofs, W. 1947: "Résolution de systèmes d'équations simultanées à un grand nombre d'inconnues," *Ossature Métallique* **12,** 187–195 (MR **10,** 70) [23b, V].

°Kolmogorov, A. N. 1946: "On the proof of the method of least squares," (Russian) *Uspekhi Matem. Nauk* (new series) **1,** No. 1, 57–70. [2a?, V].

°Kormes, Mark 1943: "Numerical solution of the boundary value problem for the potential equation by means of punched cards," *Rev. Sci. Inst.* **14,** 248–250. [10b?, 18, V].

°Kormes, Jennie P., and Kormes, Mark 1945: "Numerical solution of initial value problems by means of punched-card machines," *Rev. Sci. Inst.* **16,** 7–9. [10b?, V].

°Korn, Granino A. 1949: "Stabilization of simultaneous equation solvers," *Proc. Inst. Radio Engrs.* **37,** 1000–1002. [19A].

°Kron, Gabriel 1945: "Numerical solution of ordinary and partial differential equations by means of equivalent circuits," *J. Appl. Phys.* **16,** 172–186. [19A].

°Krüger, L. 1905: "Über die Ausgleichung von bedingten Beobachtungen in zwei Gruppen," *Veröff. des Preuss. Geod. Inst.*. Neue Folge Nr. 18. [4].

Krylov. V. I. (see Kantorovich, L. V.)

Kunz. K. S. 1947: *Lecture Notes on Numerical Analysis*, Harvard Computation Lab., part 2, chap. I. [0, 2, 2c*1*, 3, 4, 10, 11c*1*].

Kurtz, A. K. 1936: "The use of the Doolittle method in obtaining related multiple correlation coefficients," *Psychometrika* **1,** 45–51. [2a*1*].

Laderman, Jack 1948: "The square root method for solving simultaneous linear equations," *MTAC* **3,** no. 21, 13–16. [2c*1*].

Lanczos, Cornelius 1952: "Solution of systems of linear equations by minimized iterations," NBS *J. Research* **49,** 33–53. [6d].

Leavens, Dickson H. 1947: "Accuracy in the Doolittle solution," *Econometrica* **15,** 45–50. [2a*1*, 16b].

Leibler, Richard A. (see Forsythe, George E.)

°Leppert, E. L., Jr. 1947: "An application of IBM machines to the solution of the flutter determinant," *J. Aeronaut. Sci.* **14,** 171–174. [1a, 18].

Lewy, Hans: Unpublished notes on his congruence method. (Integer solutions only.) [7a].

Liebmann, H. 1918: "Die angenäherte Ermittlung harmonischer Funktionen und konformer Abbildung," *Sitzungsberichte d. Bayer. Akad. Wiss. Math.-Phys. Kl.* **47**, 385–416. [10c].

°Li͡usternik, L. A. 1947: "Remarks on the numerical solution of boundary problems for Laplace's equation and the calculation of characteristic values by the method of nets," *Trudy Mat. Inst. Steklov* **20**, 49–64 (Russian) (MR **10**, 71). [12, V].

Livingston, H. M. (see Snyder, F. E.).

Lonseth, A. T. 1942: "Systems of linear equations with coefficients subject to error," *Ann. Math. Stat.* **13**, 332–337. [16a*2*].

Lonseth, A. T. 1944: "On relative errors in systems of linear equations," *Ann. Math. Stat.* **15**, 323–325. [16a*2*].

Lonseth, A. T. 1947: "The propagation of error in linear problems," *Trans. Amer. Math. Soc.* **62**, 193–212. [16a*2*].

Lonseth, A. T. 1949: "An extension of an algorithm of Hotelling," *Proc. Berkeley Symp. Math. Stat. and Prob., 1945, 1946*, 353–357. (11c*1*, 16b].

°Magnier, André 1948: "Sur le calcul numérique des matrices," *Comptes Rendus Acad. Sci. Paris* **226,** 464–465. [V].

°Mallock, R. R. M. 1933: "An electrical calculating machine," *Proc. Roy. Soc. A* **140**, 457–483. [19A].

°Many, Abraham 1950: "An improved electrical network for determining the eigenvalues and eigenvectors of a real symmetric matrix," *Rev. Sci. Inst.* **21,** 972–974. [19A].

Margenau, Henry, and Murphy, George Moseley 1943: *The Mathematics of Physics and Chemistry*, N. Y., 480–483. [2a].

°McCann, G. D. 1949: "The California Institute of Technology electric analog computer," *MTAC* **3**, 501–511. [19A].

McPherson, J. L. 1948: "Applications of large-scale high-speed computing machines to statistical work," *MTAC* **3**, 121-126. [19].

°Meerovich, È. A. 1947: "An electrical apparatus for the solution of systems of linear algebraic equations," *Èlektriehestvo* **1947,** no. 4, 65–67. [19A].

°Mehmke, R. 1892: "Über das Seidel'sche Verfahren, um lineare Gleichungen bei einer sehr grossen Anzahl der Unbekannten durch successive Annäherung aufzulösen," *Matem. Sbornik,* Moscow **16**, 342–345. [10c, 10e].

Mehmke, R. 1930: "Praktische Lösung der Grundaufgaben über Determinanten, Matrizen, und lineare Transformationen", *Math. Annalen* **103**, 300–318 (Zb **17**, 416). [2a].

°Mehmke, R., and Nekrassof, P. A. 1892: "Auflösung eines linearen Systems von Gleichungen durch successive Annäherung," *Matem. Sbornik,* Moscow Math. Soc. **16**, 437–459. [10c?].

Menzies, G. H. (date unknown): "Normal equations resolved by approximation," *Empire Survey Review* **6** (no. 46), 474–487. [11b*1*]. (Probably c. 1941.)

Miller, J. C. P. 1947: Reply to Query, *MTAC* **2**, 375. Replies to Query by D. H. Lehmer, MTAC **1** (1944?), 203–204. [10c].

Milne, William Edmund 1949: *Numerical Calculus,* Princeton Univ. Press. [1a, 2a, 2c, 16a*2*, 20b].

Milne, W. E. 1951: "Linear equations and matrices," INA multilith. [6b*2*, 10a, 10d, 11b*1*(*A* & *F* & *I*), 12f, 17b*1*].

°Milne-Thomson, L. M. 1941: "Determinant expansions," *Math. Gaz.* **25**, 130–135. [1a*1*].

Mitchell, Herbert F., Jr. 1948: "Inversion of a matrix of order 38," *MTAC* **3**, 161–166. [2a, 19].

Montgomery, D. (see Bargmann, V).

°Morgenstern, Oskar, and Woodbury, Max A. 1950: "The stability of inverses of input-output matrices," *Econometrica* **18,** 190–192. [16a*2*].

°Morris, J. 1935: "A successive approximation process for solving simultaneous linear equations," *Aeronaut. Res. Comm.*, Report no. 1711. [10a, 10c?].

Morris, J. 1946: "An escalator process for the solution of linear simultaneous equations," *Philos. Mag.* (7) **37**, 106–120 (MR 8, 287). [2b, 4].

Morris, Joseph 1947: *The Escalator Method,* Wiley. [2b, 4, 10b, 10c].

Morris, R. H. (see Herzberger, M.)

Morrison, W. J., (see Sherman, J).

Motzkin, Theodore S. 1951: Bibliography on linear inequalities, linear programming, game strategy, economic behavior, and statistical decision functions, in preparation for probable issue by National Bureau of Standards, Los Angeles.

Motzkin, Theodore S. (see also Forsythe, George E.)

Moulton, F. R. 1913: "On the solutions of linear equations having small determinants," *Amer. Math. Monthly* **20**, 242–249. [16a*2*].

Moulton, F. R. 1936: "Significant figures in statistical constants," *Science* **84**, 574–575 [16a*2*].

°Muirhead, R. F. 1912: "A mechanism for solving equations of the nth degree," *Proc. Edinburgh Math. Soc.* **30**, 69–74. [19A].

Murphy, G. M. (see Margenau, H.)

Murphy, Francis J. 1947: *The Theory of Mathematical Machines,* N. Y. [19A].

Murray, F. J. 1949: "Linear equation solvers," *Quart. Appl. Math.* **7**, 263–274. [19A].

°Mysovskikh, I. P. 1950: "Über die Konvergenz der Methode von L. V. Kantorovich zur Lösung von Funktionalgleichungen und ihre Anwendungen," *Doklady Akad. Nauk SSSR,* N. S. **70,** 565–568 (Russian) (Zb **37,** 210) [11a(*1* & *6*)].

°Näbauer, M. 1910: "Vorrichtung zur Auflösung eines linearen Gleichungssystems," *Zeitschr. Math. Phys.* **58**, 241–246. [19A].

Nekrassof, P. A. (see Mehmke, R.)

°Neville, E. H. 1948: "Ill-conditioned sets of linear equations," *Philos. Mag.* (7), **39,** 35–48 (MR **9,** 382). [2b, 4b, 11c*1*, 16b*2*].

°Newing, S. T. 1941: "Determination of the shearing stresses in axially symmetrical shafts under torsion by finite difference methods," *Philos. Mag.* (7) **32,** 33–49. [10d?].

Nielsen, K. L., and Goldstein, L. 1947: "An algorithm for least squares," *J. Math. Phys.* **26**, 120–132. [2a*1A*].

°Nikolaeva, M. V. 1949: "On the relaxation method of Southwell (a critical survey)," *Trudy Mat. Inst. Steklov* **28,** 160–182 (Russian) (MR **12,** 539). [11b*1*].

Nörlund, N. E. 1940: "Ausgleichung nach der Methode der kleinsten Quadrate bei gruppenweiser Anordnung der Beobachtungen," *Acta Math.* **72**, 283–353. [1b].

Oldenburger, Rufus 1940b: "Convergence of Hardy Cross's balancing process," *J. Appl. Mech.* **7** A166–A170. [10c?].

°Opitz, G., and Willers, F. A. (date unknown): "Eingangs- und Rechnungsfehler bei der Auflösung eines Systems von n linearen Gleichungen," manuscript. (Mentioned by Collatz 1949.) [16].

Opler, Ascher 1951: "Monte Carlo matrix calculation with punched card machines," *MTAC* **5,** 115–120. [15b, 18].

°Oppokov, G. V. 1939: *Numerical Analysis* (Russian), Moscow-Leningrad, Oborongiz, 176 pp. [V].

°Ostrowski, Alexandre 1937a: "Sur la détermination des bornes inférieures pour une classe des déterminants," *Bull. Sci. Math.* II, **61**, 19–32 (Zb **16,** 3) (Fs **63,** 34). [16a*2*].

°Ostrowski, Alexander 1937b: "Über die Determinanten mit überwiegender Hauptdiagonale," *Comment. Math. Helvetici* **10,** 69–96 (Fs **63,** 35) (Zb **17,** 290). [16a*2*].

°Ostrowski, Alexandre 1938: "Sur l'approximation du déterminant de Fredholm par les déterminants des systèmes d'équations linéaires," *Arkiv. för Mat., Astronomi och Fysik* **26**A, no. 14, 1–15. [1a, 4].

Ostrowski, Alexandre 1950: "Sur la variation de la matrice inverse d'une matrice donnée," *Comptes Rendus Acad. Sci. Paris* **231,** 1019–1021. [16a*2*].

Ostrowski, A. M., Todd, Olga T., and Todd, John 1949: *Bibliography on Computational Aspects of Finite Matrix Theory,* N. B. S., Washington, Part I. Inversion of Matrices.

°Panov, D. I͡U. 1934: "Solution of systems of linear equations," Supplement to translation of J. Scarsborough's *Numerical Methods in Mathematical Analysis,* Moscow-Leningrad. [0a?].

°Panov, D. I͡U. 1938: *Handbook on the Numerical Solution of Partial Differential Equations,* Moscow-Leningrad, Izd. Akad. Nauk. 129 pp. [V].

°Parker, W. W. 1941: "The modern a.-c. network calculator," *Trans. Amer. Inst. Elec. Engrs.* **60,** 977–982 [19A].

°Parker, W. W. 1945: "Dual a. c. network calculator," *Electrical Engineering* **64,** 182–183. [19A].

Parkes, E. W. 1950: "Linear simultaneous equations," *Aircraft Engineering* **22,** 48, 56. [2a, 16b*2*, 22].

Petrie, George W., III. 1953: "Matrix inversion and solution, etc.," *Simultaneous Linear Equations and the Determination of Eigenvalues,* NBS, AMS29. [3a*2*, 18].

°Picchi, M. 1948: "An electrical machine for the solution of algebraic simultaneous linear equations," *Elettrotecnica* **35,** 406–410 (Italian) (SA 2860 (1949)). [19A].

°Picone, M. 1940: "Lezioni di calcolo numerico," (Roma—D. U. S. A., Città universitaria (1940–1941). [V].

Picone, M. (date unknown): "Lezioni di analisi matematica." [2c, 3b].

Pipes, Louis A. 1941: "The solution of a. c. circuit problems," *J. Appl. Phys.* **12,** 685–691 (MR **3,** 154). [20e].

Pipes, Louis A. 1946: *Applied Mathematics for Engineers and Physicists,* McGraw-Hill, chap. IV. [1a*1*].

Pipes, Louis A. 1948: "Devices for solving systems of linear algebraic equations," Division of Engineering Extension, U. C. L. A. [0, 1a*1*, 2a*1*, 4, 6b*1*, 10a, 10c, 10d, 11a (*1* & *6*), 11c*1*, 19A].

°Pirlet, Joseph 1909: "Fehleruntersuchungen bei der Berechnungen mehrfach statisch unbestimmter Systeme," Dissertation, Aachen. [16, V].

Plunkett, Robert 1950: "On the convergence of matrix iteration processes," *Quart. Appl. Math.* **7,** 419–421. [10d].

Polachek, H. 1948: "On the solution of systems of linear equations of high order," N. O. L., White Oak, Memo. NOLM–9522, 8 pp. [2a, 5, 16b*2*].

Pollaczek-Geiringer, Hilda 1928: "Zur Praxis der Lösung linearer Gleichungen in der Statik," *Zeitschr. f. Angew. Math. Mech.* **8,** 446–447 (Fs **54,** 586). [10c].

Pollackek-Geiringer, Hilda (see also Geiringer, Hilda and von Mises, R.)

Price, G. Baley 1951: "Bounds for determinants with dominant principal diagonal," *Proc. Amer. Math. Soc.* **2.** 497–502. [16a].

°Proshko, V. M. 1947: "An electrical apparatus for the solution of systems of compatible linear algebraic equations," *Trudy Mat. Inst. Steklov* **20,** 117–128 (Russian) (SA 3661 (1949)). [19A].

°Quade, W., 1947: "Auflösung linearer Gleichungen durch Matrizeniteration," *Ber. Math.-Tagung Tübingen 1946,* 123–124. [10d].

Rainsford, H. F. (date unknown): "Least-square solutions with weights," *Empire Survey Review* **7** (no. 47), 9–23. [11b*1* (*B* & *F* & *I*?)]. (Probably c. 1943.)

Redheffer, Raymond 1948: "Errors in simultaneous equations," *Quart. Appl. Math.* **6,** 342–343 (MR **10,** 152). [16a*2*].

Reich, Edgar 1949: "On the convergence of the classical iterative method of solving linear simultaneous equations," *Ann. Math. Stat.* **20,** 448–451. [10c].

°Reich, E. (date unknown): "Order of combination of arithmetical operations for minimum round-off error," Project Whirlwind Memorandum M–239. [16b*2*?, V].

Reich, Edgar 1948: "The solution of linear algebraic equations by successive approximations," M. I. T. Servomechanisms Laboratory, Memorandum M–565, 5 Aug. 1948, 36 pp. [0a, 10a, 10c, 10d, 11a (*2* & *5*), 11b*1* (*A* & *F* & *I*), 14b, 16b*2*]. (There are other memoranda with similar titles.)

Renner, H. W. 1946: "Solving simultaneous equations through the use of IBM electric punched card accounting machines," Personal Paper (write c/o IBM, Endicott, N. Y.), 6 pp. [2a, 18].

Reynolds, Walter F. 1934: *Manual of Triangulation Computation and Adjustment,* USCGS Spec. Publ. No. 138, Gov. Printing Office, 1934 (reprint 1946). [2a].

Robinson, G. (see Whittaker, E. T.)

Ricci, Lelia 1949: "Confronto fra i metodi di Banachiewicz, Roma e Volta per la risoluzione dei sistemi di equazioni algebriche lineari," *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Nat.* (8) **7,** 72–76. [2a, 2c, 3b, 22].

Rice, L. H. 1920: "Some determinant expansions," *Amer. J. Math.* **42,** 237–242. [1a*1*].

°Richardson, D. F. 1946: *Electrical Network Calculations,* D. van Nostrand. [2a].

Richardson, L. F. 1910: "The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam," *Philos. Trans.* (A) **210,** 307–357. [10d, 12e, 13c].

Robinson, R. M. 1951: "Solution of linear equations with the stencils," hectographed, Univ. of Calif. Math Dept., Berkeley, Calif., 2 pp. [7a, 18].

Rock, S. M. (see Berry, C. E.)

Roessler, Edward B. 1936: "Significant figures in statistical constants," *Science* **84,** 289–290. [16a*2*].

°Roma, Maria Sofia 1946: "Il metodo dell'ortogonalizzazione per la risoluzione numerica dei sistemi di equazioni lineari algebriche," *Ricerca Scientifica* **16,** 309–312 (MR **8,** 171). [3b].

°Roma, Maria Sofia 1947: "Il metodo dell'ortogonalizzazione per la risoluzione numerica dei sistemi di equazioni algebriche," *Pubbicazioni dell'Istituto per le Applicazioni del Calcolo,* no. 189, Roma. (Reprinted from Revista del Catasto e dei Servizi Tecnici Ereriali, no. 1 (1946), 1–12.) (MR **10,** 574). [3b].

Rosenberg, R. L. (see Stein, P.)

Rosser, J. B. 1949 (author's name not on copy): Notes on hermitian synthesis of various least-squares methods, Institute for Numerical Analysis. [11a (comprehensive), 13a (*1 & 4*), 13a (*2 & 6*), 13c*1*].

Rosser, J. B. 1950: "A general iteration scheme for solving simultaneous equations," abstract, *Bull. Amer. Math. Soc.* **56,** 176–177. [11a (comprehensive), 13c*1*, 19].

Rosser, J. Barkley 1952: "A method of computing exact inverses of matrices with integer coefficients," NBS *J. Research* **49,** 349–358. [2a*2*].

Rosser, J. Barkley 1953: "Rapidly converging iterative methods for solving linear equations," *Simultaneous Linear Equations and the Determination of Eigenvalues,* NBS AMS 29. [6d].

°Rubin, Tryggve 1926: "Ett nytt sätt att lösa normalekvationer," *Svensk Lantmäteri-Tidskrift,* Lidköping, **1.** [2c*1*].

Runge, C. 1899: "Separation und Approximation der Wurzeln," *Enc. d. Math. Wiss.* I B 3a 15, p. 448. [10c].

Runge, C., and König, H. 1924: *Vorlesungen über numerisches Rechnen,* Springer, Berlin, 183–188. [5, 10b].

Saibel, Edward 1944: "A rapid method of inversion of certain types of matrices," *J. Franklin Inst.* **237,** 197–201. [4].

Salisbury, Frank S. (see Kelley, Truman L.).

Salvadori, Mario G. 1948: *The Mathematical Solution of Engineering Problems,* McGraw-Hill, N. Y., 114–146. [0, 1a, 2a, 5, 10c, 16b*2*, 20a].

°Samssonow, K. W. 1933: "Über ein Gerät zur Lösung eines Systems von linearen Gleichungen," *Prikl. Mat. Mekh.* **2,** 309–313 (Russian, German summary) (Fs **61,** 1332). [19A].

Samuelson, Paul A. 1945: "A convergent iterative process." *J. Math. Phys.* **24,** 131–134. [21b*1*, 21b*2*].

Samuelson, P. A. 1950: "Solving linear equations by continuous substitution," abstract, *Bull. Amer. Math. Soc.* **56,** 159. [3a*2*].

Sassenfeld, H. 1951: "Ein hinreichendes Konvergenzkriterium und eine Fehlerabschätzung für die Iteration in Einzelschritten bei linearen Gleichungen," *Zeitschr. f. Angew. Math. Mech.* **31,** 92–94. [10c].

Satterthwaite, F. E. 1944: "Error control in matrix calculation," *Ann. Math. Stat.* **15,** 373–387. [2a*1A*, 2c, 16b*1*, 17b*3*].

Scarborough, James B. 1950: *Numerical Mathematical Analysis,* Johns Hopkins, 2nd edit., 38–45. [2a, 16a*2*, 16b*2*].

°Schaefer, C. 1927: *Briefwechsel zwischen Gauss und Gerling,* Otto Elsner Verlag, Berlin. [11b*1* (*B & F & I*), V].

Schmeidler, Werner 1949: *Vorträge über Determinanten und Matrizen mit Anwendungen in Physik und Technik,* Berlin, 155 pp. [1a, 10c, 10d].

Schmidt, E. 1908: "Über die Auflösung linearer Gleichungen mit unendlich vielen Unbekannten," *Rend. Circ. Mat. Palermo* **25,** 53–77. [3b].

Schmidt, R. J. 1941: "On the numerical solution of linear simultaneous equations by an iterative method," *Philos. Mag.* (7) **32,** 369–383. [6c].

Schott, Chas. A. 1855: "Solution of normal equations by indirect elimination," Report of Supt., U. S. Coast Survey, 255–264. [10b, 11b1 (*B & F & I*)].

Schott, F. (see Spangenberg, K.).

Schröder, E. 1870: "Über unendlich viele Algorithmen zur Auflösung der Gleichungen," *Math. Ann.* **2,** 317–365. [11c*3*].

Schulz, G. 1933: "Iterative Berechnung der reziproken Matrix," *Zeitschr. f. Angew. Math. Mech.* **13,** 57–59. [11c*1*, 20b].

°Schulz, H. 1938: "Elements of curve-fitting and correlation," (Reprint of Appendix C of author's *The Theory and Measurement of Demand,* Univ. of Chicago Press.) [1a, 2a, 16, V].

°Schumann, T. E. W. 1940: "The principles of a mechanical method for calculating regression equations and multiple correlation coefficients and for the solution of simultaneous linear equations," *Philos. Mag.* (7) **29,** 258–273. [19A].

Schur, I. 1917: "Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind," *J. Reine Angew. Math.* **147,** 205–232 (esp. p. 217). [4].

Schurig, O. R. (see Hazen, H. L.)

°Seidel, L. 1862: "Resultate photometrischer Messungen, etc." *Denkschriften der Münchener Akademie.* [10c].

Seidel, Ludwig 1874: "Ueber ein Verfahren, die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineäre Gleichungen überhaupt, durch successive Annäherung aufzulösen," *Abh. math.-phys. Kl.,* Bayrische Akad. Wiss., München **11** (III), 81–108. [2a, 10c, 11b*1* (C & *F* & *I*), 17b*1*, 17c].

Shanks, Daniel 1949: "An analogy between transients and mathematical sequences and some nonlinear sequence-to-sequence transforms suggested by it. Part I," NOL Memorandum 9994, 26 July 1949. [21b*1*, 21b*2*].

Shaw, F. S. 1946: "An introduction to relaxation methods (approximate methods of numerical computation)," *Council Sci. Ind. Res. (Australia) Div. of Aeronautics,* Report S. M. 78. [V].

Sherman, Jack 1951?: *Computations of Inverse Matrices by Means of IBM Machines,* Texas Co. Research Lab., Beacon, N. Y. [7b*2*, 18].

Sherman, Jack, and Morrison, Winifred J. 1949: "Adjustment of an inverse matrix corresponding to changes in the elements of a given column or of a given row of the original matrix," Abstract, *Ann. Math. Stat.* **20,** 621. [7b*2*].

Sherman, Jack, and Morrison, Winifred J. 1950: "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *Ann. Math. Stat.* **21,** 124–127 (MR **11,** 693). [16a].

Shortley, G. H., and Weller, R. 1938: "Numerical solution of Laplace's equation," *J. Appl. Phys.* **9,** 334–344. Included in Shortley-Weller-Fried 1940. [10c, 12].

Shortley, George H., Weller, Royal, and Fried, Bernard 1940: "Numerical solution of Laplace's and Poisson's equations," Ohio State Univ., *Engineering Experiment Station Bulletin* No. 107, Sept. 1940. [10c, 12].

Shortley, George (see also Flanders, Donald A.)

Shreĭder, I͡U. A. 1951: "The solution of systems of linear consistent algebraic equations," *Doklady Akad. Nauk SSSR* **76,** no. 5, 651–654 (Russian). [3b].

Smith, T. 1927: "The calculation of determinants and their minors," *Philos. Mag.* (7) **3,** 1007–1009. [1a*1*].

Snedecor, G. W. 1928: "Uses of punched card equipment in mathematics," *Amer. Math. Monthly* **35,** 161–169. [18].

Snyder, Frances E., and Livingston, Hubert M. 1949: "Coding of a Laplace boundary value problem for the UNIVAC," *MTAC* **3,** 341–350. [10c, 19].

Southwell, R. V. 1940: *Relaxation Methods in Engineering Science, a Treatise on Approximate Computation,* Oxford Univ. Press. [0c, 11b*1* (*A* & *F* & *I*), 11b*1* (*B* & *F* & *I*), 11b*1L*, 11b*2*, 20a*1*, 21a?].,

Southwell, R. V. 1946: *Relaxation Methods in Theoretical Physics,* Oxford Univ. Press. [0c, 11b*1* (*A* & *F* & *I*), 11b*1* (*B* & *F* & *I*), 11b*1L*, 11b*2*, 20a*1*, 21a].

Southwell, R. V. (see also Black, A. N.)

Spangenberg, K., Walters, G., and Schott, F. 1949: "Electrical network analyzers for the solution of electromagnetic field problems," *Proc. Inst. Radio Engrs.* **37,** 724–729 and 866–872. [19A].

Spoerl, Charles A. 1943: "A fundamental proposition in the solution of simultaneous linear equations," *Trans. Actuar. Soc. Amer.* **44,** 276–288 (MR **5,** 161). [2a?].

Spoerl, Charles A. 1944: "On solving simultaneous linear equations," *Trans. Actuar. Soc. Amer.* **45,** 18–32 and 67–69. [2a?].

Stein, Marvin L. 1952: "Gradient methods in the solution of systems of linear equations," NBS *J. Research* **48,** 407–413. [11a, 11a (*2* & *4* & *5*), 11a4].

Stein, P. 1951: "The convergence of Seidel iterants of nearly symmetric matrices," *MTAC* **5,** 237–239. [10c].

Stein, P., and Rosenberg, R. L. 1948: "On the solution of linear simultaneous equations by iteration," *J. London Math. Soc.* **23,** 111–118. [10b, 10c].

Stiefel, E. 1951: "Some special methods of relaxation technique," *Simult. Linear Equations and the Determination of Eigenvalues,* NBS AMS29. [6d].

Stiefel, E. 1952: "Über einige Methoden der Relaxationsrechnung," *Z. Angew. Math. u. Physik,* **3,** 1–33. [6d].

Swift, C. J., and Tikson, M. 1951: "Solution of differential equations by sampling methods," working paper, Div. 11, NBS, Washington. [15c*1*].

Synge, J. L. 1944: "A geometrical interpretation of the relaxation method," *Quart. Appl. Math.* **2,** 87–89. [11b*1* (*C* & *F* & *I*)].

Taussky, O. 1949: "Relations between the condition numbers of a matrix," Oscillation Sub-Committee Report 12.409 (*Aeronautical Research Council of Great Britain.*) [16a*1*, 17b*1*].

Taussky, Olga 1950: "Note on the condition of matrices," *MTAC* **4,** 111–112. [16a*1*, 17b*1*].

Taussky, Olga 1951: *Bibliography on Bounds for Characteristic Roots of Finite Matrices,* NBS Report 1162, Washington, 10 pp.

Taussky, Olga (Todd, Olga Taussky): (See also Ostrowski, A.)

Temple, G. 1939: "The general theory of relaxation methods applied to linear systems," *Proc. Roy. Soc. A* **169,** London, 476–500. [11a (*1* & *6*), 11b*1* (*B* & *F* & *I*), 17b*1*].

Thompson, E. H. (date unknown): "Least-square solutions with a calculating machine," *Empire Survey Review* **3** (no. 20), 361–364. [11b*1* (*B* & *F* & *I*)]. (Probably c. 1936.)

Tikson, M. (see Swift, C. J.)

Todd, John 1919a: "The condition of a certain matrix," *Proc. Cambridge Phil. Soc.* **46,** 116–118. [16a*1*].

Todd, John 1949b: "The condition of certain matrices. I," *Quart. J. Mech. Appl. Math.* **2,** 469–472. [16a*1*].

Todd, John 1951a: "Solution of differential equations by sampling methods. I. Experiments on a two-dimensional case using SEAC," NBS, CL 50–3, (working paper) Washington D. C. [15c*1*, 19].

Todd, John 1951b: "Matrix inversion by a Monte Carlo method," NBS, CL 50–2, (working paper), Washington, D. C. [15c*1*].

Todd, John and Todd, Olga T. (see also Ostrowski, A.)

Toeplitz, Otto (see Hellinger, Ernst).

Tolley, H. R., and Ezekiel, Mordecai 1927: "The Doolittle method for solving multiple correlation equations versus the Kelley-Salisbury 'iteration' method," *J. Amer. Stat. Assoc.* **22,** 497–500. [0c, 2a*1A*].

Tompkins, C. 1949?: "Some projection methods in calculation of some linear problems," part of Appendix I to Progress Report No. 17, Contract N6 onr—240, Project NR 047 010. [13a *(2 & 6)*].

Tryon, John G. 1951: Unpublished bibliography on analogue computation, Dept. of Engineering Physics, Cornell Univ., Ithaca, N. Y. [19A].

Tucker, Ledyard R. 1940: "A matrix multiplier," *Psychometrika* **5,** 289–294. [18].

Tuckerman, L. B. 1941: "On the mathematically significant figures in the solution of simultaneous linear equations," *Ann. Math. Stat.* **12,** 307–316. [16a*2*, 16b*2*].

Turetsky, R. 1951: "The least squares solution for a set of complex linear equations," *Quart. Appl. Math.* **9,** 108–110. [4, 20c].

Turing, A. M. 1948: "Rounding-off errors in matrix processes," *Quart. J. Mech. Appl. Math.* **1,** 287–308. [0b, 2a*1*, 2b, 2b*1*, 2c, 3a*2*, 3b, 5, 11c*1*, 16a*1*, 16a*2*, 16b*2*].

°Turton, F. J. 1945: "On the solution of the numerical simultaneous equations arising in the analysis of redundant structures," *J. Roy. Aeronaut. Soc.* **49,** 104–111 (MR **6,** 218—noted only). [V].

Ullman, Joseph 1944: "The probability of convergence of an iterative process of inverting a matrix," *Ann. Math. Stat.* **15,** 205–213. [11c*1*, 11c*3*, 16b*2*B].

Unger, H. 1951: "Orthogalisierung (Unitarisierung) von Matrizen nach E. Schmidt und ihre praktische Durchführung," *Zeitschr. f. Angew. Math. Mech.* **31,** 53–54. [3b].

van Dantzig, D. 1951: "Remarks concerning the solution of systems of linear equations," manuscript. [16a*2*].

Verzuh, Frank M. 1949: "The solution of simultaneous equations with the aid of the 602 calculating punch," *MTAC* **3,** 453–462. [2a, 2e, 18].

°Volta, Ezio 1950: "Un nuovo metodo per la risoluzione rapida di sistemi di equazioni lineari," *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Nat.* (8) **7,** 203–207 (MR **11,** 743). [3a*2*].

von Mises, R., and Pollaczek-Geiringer, Hilda 1929: "Praktische Verfahren der Gleichungsauflösung," *Zeitschr. f. Angew. Math. Mech.* **9,** 58–77 and 152–164. [10b, 10c, 10d, 17b*1*].

von Neumann, John, and Goldstine, H. H. 1947: "Numerical inverting of matrices of high order," *Bull. Amer. Math. Soc.* **53,** 1021–1099 (Sequel under Goldstine and von Neumann.) [2a*1*, 16b*2*, 19, 23b].

von Neumann, J. (see also Bargmann, V.)

°Waddell, J. A. L. 1916: *Bridge Engineering.* [10?].

Walsh, J. L. 1920: "On the solution of linear equations in infinitely many variables by successive approximations," *Amer. J. Math.* **42,** 91–96. [10d, 16a*2*].

Walters, G. (see Spangenberg, K.)

Walther, A. 1944: "Zum Determinantenverfahren von Chiò," *Zeitschr. f. Angew. Math. Mech.* **24,** 41 (MR **7,** 407). [1a*1*, 2a].

Washburn, H. W. (see Berry, C. E.)

Wasow, W. R. 1952: "A note on the inversion of matrices by random walks," *MTAC* **6,** 78–81. [15c*1*].

Waugh, F. V. 1945: "A note concerning Hotelling's method of inverting a partitioned matrix," *Ann. Math. Stat.* **16,** 216–217. [4].

Waugh, Frederick V. 1950: "Inversion of the Leontief matrix by power series," *Econometrica* **18,** 142–154. [10d, 16b*2*].

Waugh, Frederick V., and Dwyer, Paul S. 1945: "Compact computation of the inverse of a matrix," *Ann. Math. Stat.* **16,** 259–271. [2a*1A*, 2a*2*, 2c, 2d].

Wegner, U. 1950: "Numerische Methoden zur Lösung von linearen Gleichungssystemen," to come out, says Zurmühl 1950. [V].

Weller, R. (see Shortley, G. H.)

Whittaker, E. T., and Robinson, G. 1924: *The Calculus of Observations,* London, 395 pp. [1a*1*, 2a, 3a*2*, 10c, 11b*1* *(B & F & I)*].

°Wilbur, J. B. 1936: "The mechanical solution of simultaneous equations," *J. Franklin Inst.* **222,** 715–724. [19A].

Wilcox, D. E. (see Berry, C. E.)

Wilkinson, J. H. (see Fox, L.)

Willers, Fr. A. 1928: *Methoden der praktischen Analysis,* Berlin and Leipzig. (Translated 1947.) [2a, 10c, 16a*2*].

Willers, Fr. A. 1947: *Practical Analysis,* (trans. by Beyer), Dover. [2a, 10c, 16a*2*.]

Willers, F. A. (see also Opitz, G.)

Wittmeyer, Helmut 1934: *Einfluss der Änderung einer Matrix auf die Lösung des zugehörigen Gleichungssystems, sowie auf die characteristischen Zahlen und die Eigenvektoren,* Dissertation, Darmstadt. [16a*2*].

Wittmeyer, Helmut 1936a: "Einfluss der Änderung einer Matrix auf die Lösung des zugehörigen Gleichungssystems, sowie auf die characteristischen Zahlen und die Eigenvektoren," *Zeitschr. f. Angew. Math. Mech.* **16,** 287–300. Abbreviation of Wittmeyer 1934. [16a*2*].

Wittmeyer, Helmut 1936b: "Über die Lösung von linearen Gleichungssystemen durch Iteration," *Zeitschr. f. Angew. Math. Mech.* **16,** 301–310. [10a, 16b].

Woodbury, Max 1949: "The stability of input-output matrices," Mimeographed, 5 pp. Read at Boulder, Sept., 1949. [16a*2*].

Woodbury, Max A. 1950: "Inverting modified matrices," Stat. Research Group, Memorandum report 42, Princeton, N. J., 14 June 1950, 4 pp. [4, 16a*2*].

Woodbury, Max A. (see also Morgenstern, O.)

°Worch, G. 1932: "Über die zweckmässigste Art, lineare Gleichungen durch Elimination aufzulösen," *Zeitschr. f. Angew. Math. Mech.* **12,** 175–181 (Fs **58,** 582). [2a, 3a*2*].

Wren, F. L. 1937: "Neo-sylvester contractions and the solution of systems of linear equations," *Bull. Amer. Math. Soc.* **43,** 823–834. [2a].

°Wright, L. T., Jr. 1943: "The solution of simultaneous linear equations by an approximation method," Cornell Univ., College of Engineering, *Engineering Experiment Station Bull. No.* 31, 6 pp. (MR **5,** 110). [11b*1*].

Young, David M., Jr. 1950: *Iterative Methods for Solving Partial Difference Equations of Elliptic Type*, Thesis, Ph. D., Harvard Math. Dept., May 1950. (Written under G. Birkhoff.) [10b, 10c].

zur Capellen, W. Meyer 1948: *Mathematische Instrumente*, Series B, vol. 1 of *Mathematik und ihre Anwendungen in Physik und Technik*, 3rd edit., Leipzig. [0c, 19*A*].

°Zurmühl, R. 1944: "Das Eliminationsverfahren von Gauss zur Auflösung linearer Gleichungssysteme," *Ber. Inst. Prakt. Math.*, T. H. Darmstadt, Prof. Dr. A. Walther, Z. W. B. Unters. u. Mitt, Nr. 774, 11–14. [5].

Zurmühl, R. 1949: "Zur numerischen Auflösung linearer Gleichungssysteme nach dem Matrizenverfahren von Banachiewicz," *Zeitschr. f. Angew. Math. Mech.* **29,** 76–84. [2c].

Zurmühl, Rudolf 1950: *Matrizen. Eine Darstellung für Ingenieure*, xv + 427 pp. [0a, 0b, 2a, 2c, 2d, 5, 10c, 11b*1* (*B & F & I*), 16a*2*, 17a*1*, 17b*1*, 20a*1*, 20e].

Zurmühl, R. (see also Bodewig, E.)

°Zylev, V. P. 1939: "Criteria for the convergence and error estimates for the solution of systems of linear algebraic equations by means of iterations (with matrix presentation)," Moscow-Leningrad, *Sb. trudov inzh. stroit in-ta im. Kuĭbysheva* **2,** 232–245. [10?].

*Supplementary References **

Bell, William D. 1950: "Punched card techniques for the solution of simultaneous equations and other matrix operations," *Proceedings Scientific Computation Forum, 1948*, I. B. M. Corp., N. Y. 28–31. [3a*3*, 18].

Bowie, O. L. 1951: "Practical solution of simultaneous linear equations," *Quart. Appl. Math.* **8,** 369–373 (MR **12,** 538). [17c*2*, 10c].

Chancellor, Justus, Sheldon, John W., and Tatum, G. Liston 1951: "The solution of simultaneous linear equations using the IBM card-programmed electronic calculator," *Proceedings, Industrial Computation Seminar, September 1950*, I. B. M. Corp., N. Y., 57–61. [2c, 18].

Good, I. J. 1950: "On the inversion of circulant matrices," *Biometrika* **37,** 185–186 (MR **12,** 538). [23].

Grosh, L. E., Jr., and Usdin, E. 1951: "A method for evaluating determinants and inverting matrices with arbitrary polynomial elements by IBM punched card methods," *Proceedings, Industrial Computation Seminar, September 1950*, I. B. M. Corp., N. Y., 99–103. [1a, 18].

Gutshall, William D. 1951: "Practical inversion of matrices of high order," *Proceedings, Computation Seminar, December 1949*, I. B. M. Corp., N. Y., 171–173. [2a?, 4?, 16b*1*, 23b].

Kunz, Kaiser S. 1951: "Matrix methods," *Proceedings, Computation Seminar, December 1949*, I. B. M. Corp., N. Y., 37–42. [1a, 2a, 2c*1*, 10b, 10c, 20b].

Liggett, Irving C. 1951: "Two applications of the IBM card-programmed electronic calculator," *Proceedings, Industrial Computation Seminar, September 1950*, I. B. M. Corp., N. Y., 62–65. [10c, 18].

Lowe, John 1951: "Solution of simultaneous linear algebraic equations using the IBM type 604 electronic calculating punch," *Proceedings, Computation Seminar, December 1949*, I. B. M. Corp., N. Y., 54–56. [2a, 18].

Luckey, Bonalyn A. 1951: "Inversion of an alternant matrix," *Proceedings, Computation Seminar, December 1949*, I. B. M. Corp., N. Y., 43–46. [1a, 18].

Murray, Francis J. 1950: "Simultaneous linear equations," *Proceedings, Scientific Computation Forum, 1948*, I. B. M. Corp., N. Y., 105–106. [3b, 16a*1*, 17c*1*].

Porter, Randall E. 1951: "Single order reduction of a complex matrix," *Proceedings, Computation Seminar, December 1949*, I. B. M. Corp., N. Y., 138–140. [18].

Schwerdtfeger, Hans 1951?: *Bibliography on Iteration*, manuscript, reproduced by multilith at INA. (About 120 titles.)

Sheldon, John W. (see Chancellor, Justus)

Tatum, G. Liston (see Chancellor, Justus)

Usdin, E. (see Grosh, L. E., Jr.)

*Second Supplement ***

Abramov, A. A. 1950: "On a method of acceleration of iterative processes," *Doklady Akad. Nauk SSSR* **74,** 1051–1052 (Russian). (MR **12,** 861). [12f].

Bodewig, E. 1950: "Bericht über die Methoden zur numerischen Lösung von algebraischen Eigenwertproblemen," *Atti Sem. Mat. Fis. Univ. Modena* **4,** 133–193; **5** (1951), 3–39.

Brown, George W. (see Goldberg, Edwin A.)

*Most of these subjects have not been entered in section IV.

**Most of these titles have been added in proof. Some fifty more titles, mostly recent, could have been added.

Bückner, Hans 1948: "A special method of successive approximations for Fredholm integral equations," *Duke Math. J.* **15,** 197–206. [12f].

Eckert-Mauchly Division of Remington-Rand Corp. (Philadelphia, Penn.) 1951: *Matrix Algebra Programs for the UNIVAC*, ozalid of typescript, 11 pp. (no author listed).

Faddeeva, V. N. 1950: *Computational Methods of Linear Algebra*, Moscow-Leningrad, 240 pp. (Russian). Excellent. A translation is in progress at NBS, Los Angeles.

Frankel, Stanley P. 1949: *Bibliography on Computing Machines*, Hectographed by Analysis Laboratory, Calif. Inst. of Technology, 44 pp. (Author's name not included.)

°Friedrich, Konrad and Jenne, Werner 1951: *Geometrisch-anschauliche Auflösung linearer mit Nullkoeffizienten ausgestatteter Gleichungssysteme*, Deutsche Akad. Wiss. Berlin. Veröff. Geodät. Inst. Potsdam, no. 5, viii+68 pp. (MR **13,** 387).

Gavurin, M. K. 1950: "Application of polynomials of best approximation to improving the convergence of iterative processes," *Uspekhi Matem. Nauk* **5,** no. 3, 156–160 (Russian). [12f].

°Goldberg, Edwin A., and Brown, George W. 1948: "An electronic simultaneous equation solver," *J. Appl. Phys.* **19,** 339–345. [19A].

°Goursat, É. 1903: "Sur quelques développements de 1:(1−x) en séries de polynomes," *Bull. des Sciences Math.* (2) **27,** 226–232. (Fs **34,** 305).

Hestenes, Magnus R., and Stiefel, Eduard 1952: "Method of conjugate gradients for solving linear systems," NBS *J. Research* **49,** 409–436. [3b, 6d].

°Jenne, Werner (see Friedrich, Konrad).

Mitchell, Herbert F., Jr. 1950: *Solution of Matrix Equations of High Order by an Automatic Computer*, mimeographed, Eckert-Mauchly Computer Corp., 3747 Ridge Ave., Philadelphia 32, Pa., 19 pp.

Morris, J. 1935: "On a simple method for solving simultaneous linear equations by means of successive approximations," *J. Roy. Aeronaut. Soc.* **39,** 349–(unknown).

Nekrasov, P. A. 1884: "Determination of the unknowns by the method of least squares for a very large number of unknowns," *Mat. Sbornik* **12,** 189–204 (Russian). [10c].

Orden, A. 1948: *Code for Solution of Simultaneous Equations by Elimination*, M. I. T., Project Whirlwind Engineering Note E–161, 4 Nov. 1948, 32 pp. and many charts. [2a].

Ostrowski, Alexandre 1936: "Sur une transformation de la série de Liouville-Neumann," *Compt. Rend. Acad. Sci. Paris* **203,** 602–604. [11c*1*].

Ostrowski, Alexandre 1938a: "Sur quelques transformations de la série de Liouville-Neumann," *Compt. Rend. Sci. Paris* **206,** 1345–1347. [11c*3*].

°Parodi, Maurice 1951: "Sur des familles de matrices auxquelles est applicable une méthode d'itération," *Comp. Rend. Acad. Sci. Paris* **232,** 1053–1054 (MR **12,** 639).

°Pizzetti, P. 1887: "Sulla compensazione delle osservazioni secondo il metodo dei minimi quadrati, Nota I, II," *Roma, Accad. dei Lincei, Rendiconti* (4) III$_2$, 230–235 and 288–293. (Fs **19,** 213). [10c, 10e].

°Souriau, J.-M., and Bonnard, R., 1951: "Théorie des erreurs en calcul matriciel," *Recherche Aéronautique* **1951,** no. 19, 41–48 (MR **12,** 638).

°Steffensen, J. F. 1933: "Remarks on iteration," *Skandinavisk Aktuarictidskrift* **16,** 64–72. (Fs **59,** 535). [21b*1*].

Stiefel, Eduard (see Hestenes, Magnus R.)

°Terracini, Allesandro 1935: "Un procedimento per la risoluzione numerica dei sistemi di equazioni lineari," *Ricerche di Ingegneria* **3,** 40–48 (Fs **61,** 1331).

°Thiele, T. N. 1909: *Interpolationsrechnung*, Leipzig, 175 p. [Bodewig 1950 says this contains 21b*1*].

°Thomson, W. (Lord Kelvin) 1878: "On a machine for the solution of simultaneous linear equations," *Proc. Roy. Soc. London* **28,** 111–113 (Fs **10,** 111). [5].

Thurstone, L. L. 1935: *The Vectors of Mind*, Chicago, 266 pp. [2c*1*].

2. Simultaneous Systems of Equations

A. M. Ostrowski [1]

2.1. In the theory of systems of n linear equations with n unknowns, the problem of existence of olutions does not present any difficulties whatever, as soon as the determinant can be computed and is ot zero. On the other hand, the interdependence between the solutions and the coefficients can be reated in introducing different expressions corresponding to different assumptions about the norm in the ector space in question [1, 2, 3]. The theory of a system of n nonlinear equations with n unknowns ppears from the beginning as essentially more difficult, because not even the existence of solutions can e treated in a corresponding way. If the Jacobian, the analog of the determinant, is not zero, we can nly say something about the continuation of the originally given solution in a sufficiently small neighbourhood.

However, in replacing the *projective invariant*—the Jacobian—by a certain *orthogonal invariant,* a ool can be obtained which is sufficiently strong and flexible to deal with these problems and to obtain theory that presents in many respects a complete analogy with the theory in the linear case.

2.2. We will discuss the system of equations [2]

$$f_\mu(X)=f_\mu(x_1, \ldots, x_n)=0 \qquad (\mu=1, \ldots, n), \tag{1}$$

vhere X is the vector with the real components $x_1, \ldots, x_n$ and the real functions f_μ have continuous irst derivatives. The matrix of the Jacobian

$$J=\left|\frac{\partial f_\mu}{\partial x_\nu}\right| \tag{2}$$

—this matrix will be written as

$$(J)=\left(\frac{\partial f_\mu}{\partial x_\nu}\right) \tag{3}$$

—is here of particular importance. If, as will be always assumed in what follows, $J\neq 0$, we will write he inverse matrix to (J) as

$$(J^{-1})\equiv\left(\frac{\partial x_\nu}{\partial f_\mu}\right). \tag{4}$$

The symbol $\frac{\partial x_\nu}{\partial f_\mu}$ denotes here the corresponding element of J^{-1}. Put then

$$\Delta_2(X)=\left[\sum_{\mu,\nu}\left(\frac{\partial x_\nu}{\partial f_\mu}\right)^2\right]^{\frac{1}{2}}; \tag{5}$$

Δ_2 measures to a certain extent "the degree of nonvanishing" of J. Suppose now that for a vector $X_0(x^0_\nu)$ we have

$$f_\mu(X_0)=y^0_\mu, \tag{6}$$

vhere the vector $Y_0(y^0_\mu)$, that is its norm, is "small". Then we may expect that in the neighbourhood of X_0 there exists a solution of (1).

This statement can be rigorously proved in the following form:

If for a certain $D>0$ in the neighbourhood of X_0

$$|X_0-X|\leq D|Y_0| \tag{7}$$

he relation

$$\Delta_2(X)\leq D, \tag{8}$$

iolds, then there exists a solution of (1) in the neighbourhood (7).

[1] University of Basle, Switzerland.

[2] In what follows the indices $\mu, \nu, \lambda, \kappa$ will always run from 1 to n.

2.3. The theorem of section 2 can be proved as follows: Consider the system of equations

$$f_\mu(x_1, \ldots, x_n)-(1-t)f_\mu(x_1^0, \ldots, x_n^0)=0 \tag{9}$$

where t varies from 0 to 1. For sufficiently small values of t there exist solutions X_t of (9), which form a continuous arc C, beginning in X_0 and contained in (7). Then we have for the length of arc $s(t)$ along C, measured from $t=0$, by Cauchy-Schwarz inequality

$$\left(\frac{ds}{dt}\right)^2=\sum_\nu\left(\frac{dx_\nu}{dt}\right)^2=\sum_\nu\left(\sum_\mu \frac{\partial x_\nu}{\partial f_\mu} f_\mu(X_0)\right)^2 \leq \sum_\nu \sum_\mu \left(\frac{\partial x_\nu}{\partial f_\mu}\right)^2 \sum_\kappa f_\kappa(X_0)^2=\Delta_2(X)^2|Y_0|^2, \tag{10}$$

$$0 \leq \frac{ds}{dt} \leq \Delta_2(X)|Y_0| \leq D|Y_0|. \tag{11}$$

Let now τ, $0<\tau\leq 1$, be a number with the following properties:

A. For each u, $0<u<\tau$, there exists a continuous arc C_u consisting of solution points $P_t(0\leq t\leq u)$.

B. τ is the greatest number ≤ 1 with the property A.

It follows now from (11) that the length of each of the arcs C_u is $\leq uD|Y_0|<\tau D|Y_0|$, so that all points P_t on these arcs are *in the interior* of the neighborhood (7). There exists therefore a sequence t_ν, $0<t_\nu<1$, such that $t_\nu \uparrow \tau$, and that the points P_{t_ν} converge to a point P^* also situated in (7). It follows then from the continuity of f_ν in (7) that P^* is a solution of (9) corresponding to $t=\tau$. If then $\tau=1$, our theorem is proved. Suppose now $\tau<1$.

Then, since the distance $|P_0P_{t_\nu}|<\tau\, D|Y_0|$, we have $|P_0P^*|\leq\tau\, D\,|Y_0|<D|Y_0|$, P^* lies in the *interior* of (7) and by the existence theorem there exists, for a sufficiently small $\epsilon>0$, a continuous arc, C^*, through $P^*=P_\tau$ consisting of points P_t with $|t-\tau|\leq\epsilon$. But then it follows from the uniqueness theorem that a point P_{t_ν} lies on C^*. In taking then the portion of the corresponding arc C_u until P_{t_ν} and then the portion of C^* from P_{t_ν} through P^* and beyond P^*, we see that τ is not the greatest number <1 with the property A. Therefore we have $\tau=1$, and the theorem is proved.

2.4. The theorem of section 2 can be generalized in introducing a more general "norm" in the n-dimensional vector space. Let p, q be a couple of numbers, satisfying the relations

$$\frac{1}{p}+\frac{1}{q}=1, \qquad 1\leq p\leq\infty, 1\leq q\leq\infty. \tag{12}$$

Then we define as the p-norm of the vector $X(x_\nu)$

$$|X|_p\equiv\left[\sum_\nu |x_\nu|^p\right]^{\frac{1}{p}}. \tag{13}$$

The expression (5) can be generalized to

$$\Delta_p(X)=\left[\sum_{\mu,\nu}\left|\frac{\partial x_\nu}{\partial f_\mu}\right|^p\right]^{\frac{1}{p}}, \tag{14}$$

and instead of the theorem of section 2 we obtain the theorem:

If for a certain $D>0$ in the neighbourhood of X_0 defined by

$$|X_0-X|_p\leq D|Y_0|_q, \tag{15}$$

we have throughout

$$\Delta_p(X)\leq D, \tag{16}$$

there exists in this neighbourhood a solution of (1).

For $p=q=2$ we obtain the result of section 2. The two other particularly important cases are $p=\infty$, $q=1$; $p=1$, $q=\infty$.

Observe that the orthogonal invariance only holds for $p=q=2$.

2.5. In order to apply the theorems of sections 2 and 4, we would have to compute $\Delta_p(X)$, that is essentially J^{-1}, throughout a whole neighbourhood of X_0. This is usually not feasible. Practically, one will compute $\Delta_p(X_0)$ in the initial point X_0 and will try to obtain an estimate for the variation of Δ_p in a neighbourhood of X_0. This can be done systematically, if f_μ possess finite second derivatives, in using the following important inequality.

Put

$$\Delta_2^*(X)=\left[\sum_{\nu,\kappa,\lambda}\left|\frac{\partial^2 f_\nu}{\partial x_\kappa \partial x_\lambda}\right|^2\right]^{\frac{1}{2}}. \tag{17}$$

Then we have

$$\left|\operatorname{grad}\frac{1}{\Delta_2(X)}\right|\leq\Delta_2^*(X), \tag{18}$$

where grad $[1/\Delta_2(X)]$ is a vector with the components $\partial[1/\Delta_2(X)]/\partial x_\nu$. By (18) it follows from the relation

$$\int_{X_1}^{X_2}\left(\operatorname{grad}\frac{1}{\Delta_2(X)}\right)_s ds=\frac{1}{\Delta_2(X_2)}-\frac{1}{\Delta_2(X_1)}: \tag{19}$$

$$\left|\frac{1}{\Delta_2(X_2)}-\frac{1}{\Delta_2(X_1)}\right|\leq\int_{X_1}^{X_2}\Delta_2^*(X)ds. \tag{20}$$

These inequalities can be generalized in the sense of section 4, in introducing

$$\Delta_q^*(X)=\left[\sum_{\nu,\kappa,\lambda}\left|\frac{\partial^2 f_\nu}{\partial x_\kappa \partial x_\lambda}\right|^q\right]^{\frac{1}{q}}. \tag{21}$$

We obtain then

$$\left|\operatorname{grad}\frac{1}{\Delta_p(X)}\right|_q\leq\Delta_q^*(X), \tag{22}$$

and from there on

$$\left|\frac{1}{\Delta_p(X_2)}-\frac{1}{\Delta_p(X_1)}\right|\leq|X_2-X_1|_p\int_0^1\Delta_q^*(X^{(t)})dt, \tag{23}$$

where

$$X^{(t)}=X_1+t(X_2-X_1). \tag{24}$$

However, in the cases $p=1$ or $q=1$ it may become necessary to write the above inequalities in a more special form in which the *one-sided derivatives* are used.

The inequalities (18), (22) are special cases of very general inequalities valid for general determinants whose elements are differentiable functions of certain variables.

2.6. The inequalities underlying the theorem of section 2 are obtained in using the Cauchy-Schwarz inequality. Much better results can be obtained in using an orthogonal invariant of more involved kind.

If $A=(a_{\mu\nu})$ is an $n\times n$ matrix, an n-dimensional vector X is transformed by A into Y where

$$Y'=AX', \tag{25}$$

and there exist then two numbers $\Lambda\geq\lambda\geq0$, such that for any choice of X we have

$$\lambda|X|\leq|Y|\leq\Lambda|X| \tag{26}$$

and that here Λ cannot be replaced by smaller numbers and λ by greater numbers. It is well known that Λ and λ are square roots of the greatest and least fundamental value of the matrix AA^*. We will denote these expressions by $\Lambda(A)$, $\lambda(A)$. If $|A|\neq0$, we have

$$\Lambda(A^{-1})\lambda(A)=1. \tag{27}$$

Now, in the theorem of section 2, $\Delta_2(X)$ can be replaced by $\Lambda(\partial x_\nu/\partial f_\mu)$ and since this expression never exceeds, and is usually less than, $\Delta_2(X)$, we obtain in this way closer neighbourhoods of X_0.

The computation of $\Lambda(A)$ and $\lambda(A)$ presents, of course, even much greater difficulties than that of $\Delta_2(X)$. On the other hand, there exist in this case inequalities completely analogous to (18). These inequalities can be formulated in the following way.

Suppose that the vector X is given as a differentiable function of a real variable t and denote by δ the derivative with respect to t. Then we have

$$\left|\delta\Lambda\left(\frac{\partial f_\mu}{\partial x_\nu}\right)\right| \leq \Lambda\left(\delta\,\frac{\partial f_\mu}{\partial x_\nu}\right), \tag{28}$$

$$\left|\delta\lambda\left(\frac{\partial f_\mu}{\partial x_\nu}\right)\right| \leq \Lambda\left(\delta\,\frac{\partial f_\mu}{\partial x_\nu}\right). \tag{29}$$

These inequalities are special cases of two inequalities concerning general determinants and can be generalized even further in introducing the notion of distance in a Minkowski-space.

2.7. In order to give a numerical example we consider the system of equations

$$E_1 \equiv Z_1 - 0.470\,228\,201\,834\,K(Z_2) - 0.145\,308\,505\,601\,K(Z_4) - C_1 = 0,$$

$$E_2 \equiv Z_2 + 0.760\,845\,213\,036\,K(Z_1) - 0.615\,536\,707\,435\,K(Z_3) - C_2 = 0.$$

$$E_3 \equiv Z_3 + 0.615\,536\,707\,435\,K(Z_2) - 0.760\,845\,213\,036\,K(Z_4) - C_3 = 0.$$

$$E_4 \equiv Z_4 + 0.145\,308\,505\,601\,K(Z_1) + 0.470\,228\,201\,834\,K(Z_3) - C_4 = 0.$$

where

$$K(Z) = \lg\frac{72}{61-11\cos Z}; \quad C_1 = 5.430\,415\,554\,935; \quad C_2 = 5.026\,548\,245\,744;$$
$$C_3 = 4.345\,243\,969\,113; \quad C_4 = 3.769\,911\,184\,308.$$

Here the Jacobian matrix is

$$\frac{\partial(E_\mu)}{\partial(Z_\nu)} = \begin{pmatrix} 1 & -0.4702\ldots K'(Z_2) & 0 & -0.1453\ldots K'(Z_4) \\ 0.7608\ldots K'(Z_1) & 1 & -0.6155\ldots K'(Z_3) & 0 \\ 0 & 0.6155\ldots K'(Z_2) & 1 & -0.7608\ldots K'(Z_4) \\ 0.1453\ldots K'(Z_1) & 0 & 0.4702\ldots K'(Z_3) & 1 \end{pmatrix}$$

Its inverse can be easily estimated in using $|K'(Z)| \leq 11/60$.

We obtain then for the expressions introduced in section 4

$$|\Delta_2| < 2.092; \qquad |\Delta_\infty| < 1.046; \qquad |\Delta_1| < 4.183,$$

independently of the values of Z_μ. The approximate values of Z are given by

$$Z_1^0 = 5.523\,188; \quad Z_2^0 = 4.847\,788; \quad Z_3^0 = 4.244\,909; \quad Z_4^0 = 3.684\,244, \tag{30}$$

and the corresponding values of E are

$$E_1^0 = 0.0_6207537; \quad E_2^0 = 0.0_6975533; \quad E_3^0 = 0.0_6064192; \quad E_4^0 = 0.0_6289890.$$

The norms of the corresponding vectors are then

$$|E^0|_2 < 10.407\cdot10^{-7}; \qquad |E^0|_\infty < 9.756\cdot10^{-7}; \qquad |E^0|_1 < 15.372\cdot10^{-7}$$

and from the theorem of section 4 we obtain for the errors of the approximations (30):

$$|\Delta^\infty|_2 < 21.772\cdot10^{-7}; \qquad |\Delta^\infty|_\infty < 16.080\cdot10^{-7}; \qquad |\Delta^\infty|_1 < 40.810\cdot10^{-7}.$$

If we compare this with the exact solution of our system:

$Z_1^\infty=5.523\,188\,291\ldots$; $Z_2^\infty=4.847\,788\,930\ldots$; $Z_3^\infty=4.244\,908\,849\ldots$; $Z_4^\infty=3.684\,244\,294\ldots$,

where the true values of $|\Delta^\infty|_p$ are

$$|\Delta^\infty|_2=10.29\cdot 10^{-7};\qquad |\Delta^\infty|_\infty=9.3\cdot 10^{-7};\qquad |\Delta^\infty|_1=16.66\cdot 10^{-7},$$

we see that our estimates are of the correct order of magnitude.

2.8. The expressions introduced in the above discussion of the system (1) and the properties of these expressions permit us now to give a very simple proof for the convergence of the Newton-Raphson method and to deduce very close estimates for the error implied.

We start with an initial point A_0 and define an infinite sequence of points

$$A_\kappa(a_1^{(\kappa)},\ldots,a_n^{(\kappa)})\qquad(\kappa=0,1,\ldots)\tag{31}$$

by means of the following recurrent procedure:

Let

$$Y_\kappa(y_1^{(\kappa)},\ldots,y_n^{(\kappa)}).\qquad y_\nu^{(\kappa)}=f_\nu(A_\kappa)\tag{32}$$

be the vector formed by the values of the f_ν in A_κ. Define then the vector

$$\eta_\kappa(h_1^{(\kappa)},\ldots,h_n^{(\kappa)})\tag{33}$$

from the system of linear equations

$$\sum_\mu h_\mu^{(k)}f_{\nu x_\mu}(A_\kappa)+y_\nu^{(\kappa)}=0\qquad(\nu=1,\ldots,n).\tag{34}$$

Then we have

$$A_{\kappa+1}=A_\kappa+\eta_\kappa.\tag{35}$$

In order to ensure the possibility of this procedure being continued indefinitely, we make the following assumptions: For a certain positive number M and a couple of numbers p, q satisfying (12) we suppose that throughout the neighbourhood U of A_0, defined by

$$(U)|X-A_0|_p\leq 2|\eta_0|_p,\tag{36}$$

we have for $p=q=2$

$$\Delta_2^{\bullet}(X)\leq M\qquad(X\prec U)\tag{37}$$

and otherwise

$$\left[\sum_{\kappa,\lambda,\nu}\left|\frac{\partial^2 f_\nu(X_\nu)}{\partial x_\kappa\partial x_\lambda}\right|^q\right]^{\frac{1}{q}}\leq M\qquad(X_1,\ldots,X_n\prec U).\tag{38}$$

If we have

$$\theta_0\equiv 2M|\eta_0|_p\Delta_p(A_0)\leq 1,\tag{39}$$

then all points A_κ defined by our procedure remain inside U and converge to a point A in U satisfying the equations (1). We have then in particular

$$|\eta_\kappa|_p\leq\frac{\theta_0^{2^\kappa}}{2^\kappa M\Delta_p(A_0)}\tag{40}$$

and, putting

$$\theta_\kappa=2M|\eta_\kappa|_p\Delta_p(A_\kappa):\tag{41}$$

$$\theta_{\kappa+1}\leq\frac{\theta_\kappa^2}{2-\theta_\kappa},\tag{42}$$

$$|Y_{\kappa+1}|_q\leq\tfrac{1}{2}M|\eta_\kappa|_p^2.\tag{43}$$

The estimates obtained can be sharpened in so far as the factor 2 in (36) and (39) can be replaced by a smaller one, if $2M|Y_0|_q\Delta_p^2(A_\kappa)<1$. These estimates correspond exactly to those obtained by the author for $n=1$. For $n=2$, F. A. Willers gave in 1928 conditions for the convergence of the Newton-Raphson procedure. However, these conditions involved the third derivatives of f_μ. In 1936 the author gave, again for $n=2$, conditions for the convergence implying only the second derivatives and these conditions were then generalized to the general n in the Braunschweig Thesis of K. Bussmann. However, these conditions gave only rough estimates, since the expressions Δ_p, $\overset{*}{\Delta}_2$ were not used.

On the other hand, a part of our results of section 8 has been given and considerably generalized in [4, 5, 6], but I have not been as yet able to study these papers.

It may finally be mentioned that a part of our results was published in [7, 8, 9], and an exposition of the bulk of theory will be given in a monograph that is to appear in the series of Cahiers Scientifiques (Gauthiers Villars, Paris).

[1] O. Blumenthal, Uber die Genauigkeit der Wurzeln linearer Gleichungen, Z. Math. Phy. **62**, 359–362 (1914).

[2] A. M. Ostrowski, Sur la détermination des bornes inférieures pour une classe des déterminants. Bul. Sc. Math. [2 **61**, 1–14 (1937).

[3] A. T. Lonseth, The propagation of error in linear problems, Trans. Am. Math. Soc. **62**, 193–212 (1947).

[4] L. V. Kantorovich, Functional Analysis and Applied Mathematics, Uspekhi Mat. Nauk, N. S. **3**, 89–185 (1948).

[5] L. V. Kantorovich, On Newton's method, Trudy Mat. Inst. Steklov. **28**, 104–144 (1949).

[6] J. P. Mysovskih, On convergence of Newton's method, Trudy Mat. Inst. Steklov **28**, 145–147 (1949).

[7] A. M. Ostrowski, Sur la variation de la matrice inverse d'une matrice donnée, Compt. rend. **231**, 1019–1021 (1950).

[8] A. M. Ostrowski, Un théorème d'existence pour les systèmes d'équations, Compt. rend. **231**, 1114–1116 (1950).

[9] A. M. Ostrowski, Un nouveau théorème d'existence pour les systèmes d'équations, Compt. rend. **232**, 786–788 (1951)

3. The Geometry of Some Iterative Methods of Solving Linear Systems

Alston S. Householder [1]

The purpose of this paper is to give a simple geometric derivation of some standard iterative methods for solving linear systems, and to obtain these as particular cases of a fairly general family of methods. By iterative I mean infinitely iterative, and hence belonging to Forsythe's class III.[2] My purpose is essentially the reverse of that of Dr. Forsythe, since I am trying to find the common identity rather than the points of distinction. The point of view taken here is very similar to that found in a set of notes, not signed, which were distributed by the NBS Institute for Numerical Analysis about two years ago. I hope the differences are sufficient to justify this exposition.

In order that each symbol may be easily recognized for what it represents, I shall use always small Greek letters for scalars; small roman letters for column vectors; and capitals for matrices. The space of a matrix will be the space of its column vectors. When we happen to be interested in the rows, the matrix will be designated as a transpose. A matrix A will have columns a_i, and this alphabetic association will be followed in general, except that the identity matrix I will have columns e_i.

In a space whose metric is represented by the positive definite matrix G, a projector P is a symmetric matrix of the form

$$P=U(U^TGU)^{-1}U^T,$$

with the understanding, of course, that the columns of U are linearly independent. If the columns of U are taken to be the contravariant representations of geometric vectors in some subspace, and if r is the covariant representation of any vector, then Pr is a linear combination of columns of U and hence the contravariant representation of a vector in the subspace of U. The columns of GU are the covariant representations of the same vectors of which the columns of U are the contravariant representations. Since $PGU=U$, it follows that P projects any vector in the space of U into itself. Since, further, $PGP=P$, $P(I-GP)=0$, it follows that for any covariant r, Pr is the contravariant, GPr the covariant representation of the projection, $(I-GP)r$ the covariant representation of the residual, and the projection and residual are orthogonal. Now suppose the equations to be solved are $Ax=y$, where A is nonsingular. For the iterative procedure one selects a rule for obtaining a sequence of spaces represented by the matrices U_0, U_1, U_2, . . .; then if x_α is any approximation to the solution x, one projects the residual, $s_\alpha=x-x_\alpha$ orthogonally upon the space of U_α, adjoining the projection to x_α to obtain $x_{\alpha+1}$. Whether the matrices U_α are known explicitly in advance, or whether each is selected only at the time it is to be used in accordance with some criterion that cannot be applied in advance, makes very little difference in principle. Clearly the norm of $x_{\alpha+1}$ cannot exceed that of x_α. It seems clear intuitively how convergence can be assured, and in any case I shall not stop for convergence theorems here.

If the matrix A is positive definite it is natural to let it provide the metric for the space. The equations then define the contravariant representation x of the vector whose covariant representation is the known vector y. Any approximation x_α has associated with it a covariant vector, $y_\alpha=Ax_\alpha$, which deviates by a calculable amount, $r_\alpha=y-y_\alpha$, from the required one. The vector r_α, which is computable, is the covariant representation of the deviation, while s_α is the (unknown) contravariant representation.

When we associate a matrix U_α of contravariant vectors, we can define a projector

$$P_\alpha=U_\alpha(U_\alpha^TAU_\alpha)^{-1}U_\alpha^T,$$

and then $P_\alpha r_\alpha$ is the contravariant representation of the projection which is to be adjoined to x_α:

$$x_{\alpha+1}=x_\alpha+P_\alpha r_\alpha.$$

[1] Oak Ridge National Laboratory.

[2] G. E. Forsythe, Tentative classification of methods and bibliography on solving systems of linear equations. See paper 1, page 1.

Each projection involves the inversion $(U_\alpha^T A U_\alpha)^{-1}$ of a matrix whose order is equal to the dimensionality of the space U_α upon which the projection is to be made. As the whole purpose of the method is to avoid an explicit inversion, one ordinarily takes U_α to be a single vector u_α so as to have only a scalar to invert. In particular instances, however, the matrix A might have principal minors of low order which are easily inverted. If so, a matrix U_α (or each of an infinite subset of them) can be made up of columns e_i of the identity so that the product $U^T AU$ will be such a principal minor. In general, of course, the higher the dimensionality of the space upon which the projection is made, the larger the projection and hence the greater the gain by that particular step. It is naturally possible to choose other matrices U_α yielding matrices of low order that are easily inverted, but hardly worth while to spend much time looking for them.

Another case where it might be advantageous to use matrices U_α with more than one column arises when a special purpose machine is available for inverting matrices of limited size. The Oak Ridge linear equation solver will solve systems of order 300. No larger systems have been attempted so far, but it is possible, in principle, to solve, say, a system of order 600 by inverting two matrices of order 300 each in order to project the residuals alternately upon two 300-dimensional subspaces.

The simplest, and most common, procedure is to take each U_α to be a single vector U_α. The method of the steepest descent takes $U_\alpha = r_\alpha$. That is, one would like to project upon the residual itself, in which case the problem would be completely solved in one step. But this requires the contravariant representation s_α, which we do not have. So we take r_α, the *co*variant representation, and project upon the vector for which this is the *contra*variant representation. If r_α happened to be a proper vector of A, this would still bring us out at one step, but unfortunately it generally is not.

The more common choice is to make the U_α equal to the e_i taken in some sequence. If taken in strict cyclic order we have the method of Seidel, the only feasible method for machine computation. For hand work, one can examine r_α each time and select that e_i upon which r_α has the largest projection as in the method of relaxation. If $\rho_{\alpha i}$ is the ith element in r_α, then the vector $e_i \rho_{\alpha i}/\alpha_{ii}$ is the projection upon e_i, and as this is the contravariant representation the squared length is $\alpha_{ii}(\rho_{\alpha i}/\alpha_{ii})^2 = \rho_{\alpha i}^2/\alpha_{ii}$. The optimal choice of e_i is determined by the maximum quotient of this type.

If the matrix A is not positive definite, nor, perhaps, even symmetric, the equations may be regarded in either of two ways, as representing a system of hyperplanes, or as defining the resolution of a known vector y along the column vectors of A. In the latter interpretation the elements of x are the required multipliers.

In this case it is natural to take the ordinary metric, $G = I$, in which case the projectors are

$$P_\alpha = U_\alpha (U_\alpha^T U_\alpha)^{-1} U_\alpha^T,$$

and the projection of r_α upon U_α is

$$P_\alpha r_\alpha = U_\alpha (U_\alpha^T U_\alpha)^{-1} U_\alpha^T r_\alpha.$$

However, the projection $P_\alpha r_\alpha$ is to be adjoined to $y_\alpha = Ax_\alpha$, whereas x_α itself is to be corrected by the vector $A^{-1}P_\alpha r_\alpha$. Therefore, to make this feasible we must set $U_\alpha = AV_\alpha$ and write

$$P_\alpha = AV_\alpha (V_\alpha^T A^T A V_\alpha)^{-1} V_\alpha^T A^T.$$

Hence

$$x_{\alpha+1} = x_\alpha + V_\alpha (V_\alpha^T A^T A V_\alpha)^{-1} V_\alpha^T A^T r_\alpha.$$

One chooses, thus, the V_α, and only indirectly the U_α.

The appearance of the product $A^T A$ relates this method to the result of replacing the equations by the equivalent set $A^T Ax = A^T y$ and proceeding as before with a matrix that is now positive definite. However, the complete product $A^T A$ is not required, but only products of the type $(AV_\alpha)^T (AV_\alpha)$. In particular, if each V_α is a single e_i, then one requires only the $a_i^T a_i$ to be inverted. These can be computed in advance and used in rotation to give

$$x_{\alpha+1} = x_\alpha + e_i (a_i^T a_i)^{-1} (a_i^T r_\alpha).$$

Any selection of the e_i as in the method of relaxation is ruled out since to make a selection one would have to compute every $a_i^T r_\alpha$ and compare magnitudes, whereas having computed one of these scalar products most of the work has been done for that particular projection. Note that since the projection is made upon e_i only a single element of x_α is changed, but that no single element of the residual $r_{\alpha+1}$ is necessarily eliminated.

As was indicated in the INA notes, this method applies directly to the equations of least squares. These equations have the more general form $Ax=y+e$, where e is required to be orthogonal to the space of A. The usual procedure is to multiply by A^T, which eliminates e and yields a system $A^TAx=A^Ty$ with positive definite matrix. However if one projects upon subspaces represented by $U_\alpha=AV_\alpha$, then one is projecting upon subspaces of the space of A, and the residual is always orthogonal to A. It is clear intuitively that except for very special choices of the matrices V_α the sequence y_α should approach a limit, which will necessarily lie in the space of A, and that the residuals $y-y_\alpha$ will approach a vector orthogonal to this space. However, again, I shall not stop to give a proof.

To consider the hyperplane interpretation let the equations be written in the form $A^Tx=z$. Then each column a_i of A represents the normal to one of the hyperplanes. The projection of the deviation s_α upon U_α is

$$P_\alpha s_\alpha=U_\alpha(U_\alpha^TU_\alpha)^{-1}U_\alpha^Ts_\alpha,$$

but since it is

$$t_\alpha=z-z_\alpha=A^Ts_\alpha$$

that is known, and not s_α itself, we must, as before, let $U_\alpha=AV_\alpha$ so that

$$P_\alpha s_\alpha=AV_\alpha(V_\alpha^TA^TAV_\alpha)^{-1}V_\alpha^Tt_\alpha,$$

and

$$x_{\alpha+1}=x_\alpha+AV_\alpha(V_\alpha^TA^TAV_\alpha)^{-1}V_\alpha^Tt_\alpha.$$

If for V_α one takes a vector e_i, the result is

$$x_{\alpha+1}=x_\alpha+a_i(a_i^Ta_i)^{-1}\tau_{\alpha i}$$

Again one needs then only the diagonal elements of A^TA, and not the entire matrix. To follow the method of relaxation, one should choose that i for which the projection is maximized, and the square of the projection is $\tau_{\alpha i}^2/(a_i^Ta_i)$. Whatever i is chosen, the projection is along the vector a_i so that in general every element of x_α is affected. However, since the new residual, $x-x_{\alpha+1}$, is orthogonal to a_i, the vector $x_{\alpha+1}$ taken from the origin terminates on the ith hyperplane, so that by the projection the ith residual element is caused to vanish. This is to be contrasted with the former case where no residual element will necessarily vanish but only one element of x_α requires modification.

The method of steepest descent takes for V_α the single vector t_α. The reason is that if one defines the function

$$\varphi(x)=(z^T-x^TA)(z-A^Tx)=z^Tz-2x^TAz+x^TAA^Tx,$$

then $\varphi(x)$ takes on the minimum value zero at the point x which satisfies the equations, and has the gradient $-2At_\alpha$ at the point x_α. Hence taking t_α for V_α is equivalent to taking At_α for U_α, that is a vector in the direction of the gradient, in which direction the variation of the function is most rapid. Equivalent to the original set of equations is the system $AA^Tx=y$, where $y=Az$. The gradient of $\varphi(x)$ at x_α is also expressible as $-2r_\alpha$, where $r_\alpha=y-AA^Tx_\alpha$, which accounts for the choice of vector upon which one projects in applying the method of steepest descent in the ordinary case where the matrix is positive definite.

4. Solutions of Linear Systems of Equations on a Relay Machine

Carl-Erik Fröberg [1]

Characteristics of the Bark

The BARK has been described in some detail elsewhere,[2] but for concreteness a brief review will be given here. Numbers are represented in the form $\pm 2^{\pm p} \cdot q$, where p is a six-digit binary number, and q is a 24-digit binary number. Transfer within the machine is accomplished along three busses, each of which contains 32 channels. In its first stage of development, the machine had 50 relay registers, 100 constant registers, and 840 order points; by now the corresponding figures are 100, 200, and 1,200. The arithmetic circuits consisting of an adder and a multiplier, carry out transfer, addition, and multiplication, and division is accomplished through an iteration process. Further, there are special circuits that can handle other elementary operations (for example, transfer of exponents, of the numerical part only, of the fractional part of a number, and right or left shift). Extraction of the integral part of a number needs no special order, as this can be done by adding an "integral zero", that is, zero with the exponent +24.

Orders are written in the form: n A op signs B C D, where A and B are the addresses of the two numbers which shall be combined by "op"; C is the address where the result should be stored and D the next order to be executed. Normally the order n is followed by n+1; unconditional jumps are made by means of plugged connections on the sequence panel, while conditional jumps also need selectors, consisting of relay pyramid circuits.

Input and output devices make use of standard teletype equipment. There are five tape readers, five tape punches, and one page printer available. This equipment is also used as an extra external memory, although with limited flexibility.

The operation-times are approximately:

Transfer	100 ms.
Addition	150 ms.
Multiplication	250 ms.

Methods for Solution of Linear Systems

There exist several standard methods for solution of linear systems: direct, iterative, and statistical methods. Out of these, only direct methods have been used on the BARK, and some of them will be discussed briefly.

For concreteness, we introduce the equation .

$$Ax = b, \tag{1}$$

where A is a square matrix of order n and x and b column vectors. Now it turns out that the amount of work needed for evaluation of the inverse matrix is only a factor 2 to 3 times that for solution of the equation for one special column vector b. Because of the great advantage in knowing the inverse, for example, for the solution of the equation for any vector b or for the construction of more accurate solutions from a first approximation, one very often does this extra amount of work.

The best-known direct methods are those of Gauss, Jordan, and Cholesky, but there are several other methods that can also be used.[3,4] The method of Gauss consists in forming an upper triangular

[1] University of Lund, Sweden.

[2] Kjellberg-Neovius, The BARK, A Swedish general-purpose relay computer, MTAC [33] **V**, 29 (1951).

[3] L. Fox, H. D. Huskey, and J. H. Wilkinson, Notes on the solution of algebraic linear simultaneous equations, Quart. J. Mech. and Applied Math. **1**, 149 (1948).

[4] A. M. Turing, Rounding-off errors in matrix processes, Quart. J. Mech. and Applied Math. **1**, 287 (1948).

matrix C by left-multiplication by a sequence of matrices $J_1 \ldots J_{n-1}$, where the matrix J_k differs from the unit matrix only in the kth column. Then we have

$$C = J_{n-1} \ldots J_1 A, \tag{2}$$

and the equation has taken the form:

$$Cx = d. \tag{3}$$

The unknown quantities x can now easily be obtained one after another by a simple substitution process. The inversion of C proceeds analogously, and then the inverse of A can be found directly:

$$A^{-1} = C^{-1} J_{n-1} \ldots J_1. \tag{4}$$

The method of Jordan is rather similar, but here, instead of the triangular matrix, a diagonal matrix is formed by elimination. The Cholesky method is originally applicable to symmetric matrices only and consists in expressing the matrix A in the form

$$A = LL', \tag{5}$$

where L is a lower triangular matrix and L' its transpose. Now it can be proved that A being a nonsymmetric matrix, the principal minors of which are nonsingular, there exists a unique resolution [5]

$$A = LDU, \tag{6}$$

where L and U are unit lower and unit upper triangular matrices and D a diagonal matrix. In all these cases, the inversion of a matrix is reduced to inversion of a triangular matrix, which is performed by iterated substitutions.

The Automatic Computation Aspect

There are certain limitations in the choice of computational method. Due to the small capacity of the internal memory in a relay machine, it is generally impossible to store the whole matrix there. Instead, it is put on tape, row by row, or column by column. Because punching is rather time-consuming, it is desirable that storing of intermediate results can take place in the internal memory, which means that only one single row or column can be handled simultaneously. All elimination methods being linear, it is evident that the computations can be checked by taking the sum of all elements and performing the same operations on it. This procedure takes very little time and only a few extra orders.

The operation times on a relay machine being rather long, it is desirable to choose a method where the total amount of operations is as small as possible. For such a method the number of orders will be larger than for other methods, but this only means a little extra work in coding the problem. On the other hand, by using a fast electronic computer, one can afford to use more operations, that is, longer time, in order to simplify the programming. For this reason the direct methods seem to be best adapted to a relay machine, while iterative or statistical methods might be more suitable for a fast computing machine.

On the BARK several linear systems (origin: surveying) have been solved by Gauss' and Jordan's methods. The method of Cholesky offers no special advantages here. It turns out that the time consumed when using the Gauss method for inversion of an unsymmetrical matrix is only 75 to 90 percent of that required when using the Jordan method. On the other hand, the Gauss method has the disadvantage that the inversion must take place in three steps: (a) Forming the triangular matrix C, (2); (b) inversion of C; (c) forming A^{-1}, (4); whereas the Jordan method works in two steps. Thus it is obvious that the Gauss method takes more memory space and more orders (116 compared with 87)

[5] J. von Neumann and H. H. Goldstine, Numerical inverting of matrices of high order, Bul. Am. Math. Soc. **53**, 1021 (1947).

than the Jordan method. So we can make the conclusion that, when enough memory space is available (the order of the matrix being not too large), the Gauss method should be preferred to the Jordan method.

Lastly, it should be mentioned that there exist some very special cases when the solution ought to be made by hand, namely, (a) when there are only very few elements outside the diagonal in A; and (b) when just two to three correct figures are required.

Results

As mentioned above, solution of linear systems of equations has been accomplished with the Gauss and Jordan elimination methods. Matrices of orders $n=8$, 14, 20, and 28 have been treated. Inversion of matrices of orders $n=8$ and 20 has been performed with Jordan's method.[6] The coding of the problems is straightforward; because of the floating binary point, positioning for size is unnecessary. The matrix elements are translated to octal system and put on the tape; conversion from octal to binary form is trivial.

The time consumed appears in the following table (hours):

n	Solution		Inversion
	Unsymmetrical (Jordan)	Symmetrical (Gauss)	Unsymmetrical (Jordan)
8	0. 6	0. 4	1. 1
20	3. 5	2. 1	7. 5

These results were obtained when the memory capacity was only 50 places; by now 100 memory places being available the times can be reduced considerably. No definite figures have been obtained so far, but the following formulas will give an idea of the time-consumption expected (in seconds).

Solution:

$T=0.8n^3+2.0n^2$	Unsymm.	Jordan.
$T=0.6n^3+2.0n^2$	Symm.	Jordan.
$T=0.5n^3+2.7n^2$	Unsymm.	Gauss.
$T=0.2n^3+2.7n^2$	Symm.	Gauss.

Inversion:

$T=1.9n^3+3.9n^2$	Unsymm.	Jordan.
$T=1.2n^3+3.9n^2$	Symm.	Jordan.
$T=1.4n^3+5.6n^2$	Unsymm.	Gauss.
$T=0.7n^3+5.6n^2$	Symm.	Gauss.

For example, the time for inversion of a 20×20 unsymmetrical matrix will go down from 7.5 to 4.6 hours with the Jordan method and to 3.7 hours with the Gauss method.

Most of the matrices treated have been "well-conditioned",[7] and in these cases the results have

[6] All these computations were performed under the direction of O. Karlqvist.

[7] See footnote 4.

been correct to about six digits (the machine itself works with slightly more than seven decimal digits). As an example, the following check is given:

$$10^7\ (AA^{-1}-E) \qquad (n=8)$$

$$\begin{array}{cccccccc}
-3 & +2 & 0 & -1 & 0 & -1 & 0 & 0 \\
 & -5 & +1 & 0 & 0 & 0 & -1 & 0 \\
 & & 0 & -3 & 0 & -6 & -3 & 0 \\
 & & & +1 & +2 & -4 & +3 & +1 \\
 & & & & -1 & +3 & +2 & +1 \\
 & & & & & -9 & -2 & -1 \\
 & & & & & & -7 & -3 \\
 & & & & & & & -3
\end{array}$$

where A is a symmetric 8×8 matrix with a clear diagonal dominance. The "condition-number" of A (defined as $1/n\cdot N(A)\cdot N(A^{-1})$) was 5.5.

On the other hand also, some rather ill-conditioned systems have been treated; in these cases the accuracy decreased to about three correct (decimal) figures.

Nevertheless, these results seem to indicate that the figures given by v. Neumann-Goldstine[8] normally are too pessimistic, and they rather point in favor of their new results.[9]

[8] See footnote 5.

[9] John v. Neumann and H. H. Goldstine, Numerical inverting of matrices of high order, II. Proc. Am. Math. Soc. 2, 188 (1951).

5. Some Special Methods of Relaxation Technique

Eduard Stiefel [1]

5.1. The principle of relaxation methods is to replace the direct solution of a symmetrical system of linear equations

$$\sum_{k=1}^{n} a_{ik}u_k+l_i=0, \qquad a_{ik}=a_{ki}, \qquad i=1, 2, \ldots, n \tag{1}$$

by the problem of minimizing the quadratic function

$$F=\frac{1}{2}\sum_{i,k} a_{ik}u_iu_k+\sum_i l_iu_i. \tag{2}$$

It will be assumed that the quadratic form

$$\sum_{i,k} a_{ik}u_iu_k \tag{3}$$

is positive definite. Introducing the n-dimensional vectors

$$\boldsymbol{u}=(u_1,u_2, \ldots, u_n), \qquad \boldsymbol{l}=(l_1, l_2, \ldots, l_n),$$

which may be considered as points in a space R^n, and also the matrix operator $D\equiv(a_{ik})$, the system of equations may be written as

$$D\boldsymbol{u}+\boldsymbol{l}=0 \tag{4}$$

and the quadratic function as

$$F(u)=\frac{1}{2}\,\boldsymbol{u}\cdot D\boldsymbol{u}+\boldsymbol{l}\cdot\boldsymbol{u}. \tag{5}$$

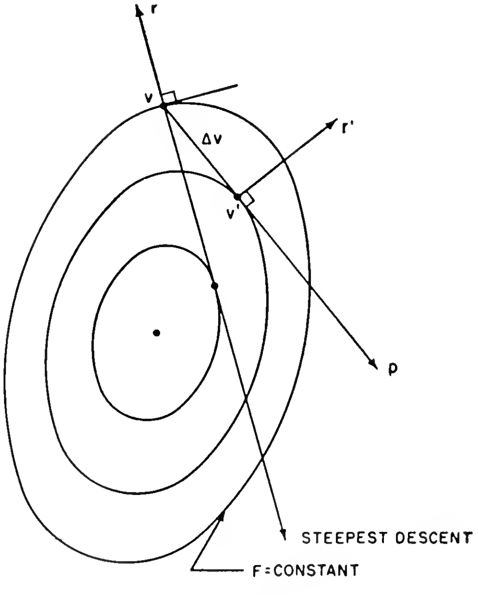


FIGURE 5.1.

[1] Professor of Applied Mathematics, Swiss Federal Institute of Technology, Zurich, Switzerland.

If the exact solution u be replaced by a trial solution $\boldsymbol{v}=(v_1, v_2, \ldots, v_n)$, then (4) must be changed to

$$D\boldsymbol{v}+\boldsymbol{l}=\boldsymbol{r}. \tag{6}$$

$\boldsymbol{r}=(r_1, r_2, \ldots, r_n)$ is called the residue vector, and its components are the residues. Taking the quadratic function $F(v)$ of the trial vector $\boldsymbol{v}$, we note that

$$\boldsymbol{r}=\text{grad } F(\boldsymbol{v}). \tag{7}$$

The idea now is to make the residues successively smaller, or, in other words, to minimize $F(\boldsymbol{v})$. The step-by-step procedure by which this is done may be best understood by looking at figure 5.1. It shows a two-dimensional cross section through R^n, which has been passed through the trial point $\boldsymbol{v}$ and the corresponding residue vector $\boldsymbol{r}$. The intersections of the surfaces $F=$constant with this cross section are a set of concentric ellipses (contour lines). As a consequence of (7), the residue vector $\boldsymbol{r}$ is normal to the ellipse passing through $\boldsymbol{v}$. One step in the computation now consists in choosing a suitable direction vector $\boldsymbol{p}$ and moving out from the trial point in the direction of $\boldsymbol{p}$ until $F(\boldsymbol{v})$ reaches a minimum value. This will be the case at some point $\boldsymbol{v}'$ at which the vector $\boldsymbol{p}$ is tangent to one of the contour lines. $\boldsymbol{v}'$ is our new trial point, and the new residue vector $\boldsymbol{r}'$ at this point will have to be perpendicular to $\boldsymbol{p}$.

Analytically we have

$$\boldsymbol{v}'=\boldsymbol{v}+\lambda\boldsymbol{p}, \tag{8}$$

where λ is a still undetermined scalar. From (6) it follows that

$$\boldsymbol{r}=D\boldsymbol{v}+\boldsymbol{l}$$

$$\boldsymbol{r}'=D\boldsymbol{v}'+\boldsymbol{l}$$

$$\boldsymbol{r}'-\boldsymbol{r}=D(\boldsymbol{v}'-\boldsymbol{v})=\lambda D\boldsymbol{p}$$

or

$$\boldsymbol{r}'=\boldsymbol{r}+\lambda D\boldsymbol{p}. \tag{9}$$

Writing the scalar product of two vectors as $\boldsymbol{a}\cdot\boldsymbol{b}$, the condition that $\boldsymbol{r}'$ be orthogonal to $\boldsymbol{p}$ becomes $\boldsymbol{r}'\cdot\boldsymbol{p}=0$, and hence

$$(\boldsymbol{r}+\lambda D\boldsymbol{p})\cdot\boldsymbol{p}=\boldsymbol{r}\cdot\boldsymbol{p}+\lambda D\boldsymbol{p}\cdot\boldsymbol{p}=0$$

or

$$\lambda=-\frac{\boldsymbol{r}\cdot\boldsymbol{p}}{\boldsymbol{p}\cdot D\boldsymbol{p}}. \tag{10}$$

After determining λ from this formula, the new trial point is given by (8) and the new residue by (9).

5.2. The various relaxation methods differ from each other in the way of choosing the direction vector $\boldsymbol{p}$. The simplest procedure is to take $\boldsymbol{p}$ parallel to one of the coordinate axes. This is essentially what is done also in Southwell's[2] "block" and "group" relaxations, which have been studied in this connection by the author.[3] This procedure has the advantage, that the computer is left considerable freedom in choosing the direction vector to suit the residue situation. On the other hand, with the use of automatic calculators it is necessary for $\boldsymbol{p}$ to be determined by some definite rule. This is most commonly done by using the method of steepest descent, in which $\boldsymbol{p}$ is chosen as $\boldsymbol{p}=-\boldsymbol{r}$; that is, we look for the minimum along the normal to the contour line at the point v. From our general formulas we have

$$v'=\boldsymbol{v}-\lambda\boldsymbol{r}, \quad \boldsymbol{r}'=\boldsymbol{r}-\lambda D\boldsymbol{r} \tag{11}$$

with

$$\lambda=\frac{\boldsymbol{r}\cdot\boldsymbol{r}}{\boldsymbol{r}\cdot D\boldsymbol{r}}. \tag{12}$$

[2] R. V. Southwell, Relaxation methods in engineering science (Oxford, 1946); Relaxation methods in theoretical physics (Oxford, 1946).

[3] E. Stiefel, Ueber einige Methoden der Relaxationsrechnung, Z. angew. Math. Phys. (Jan. 1952)

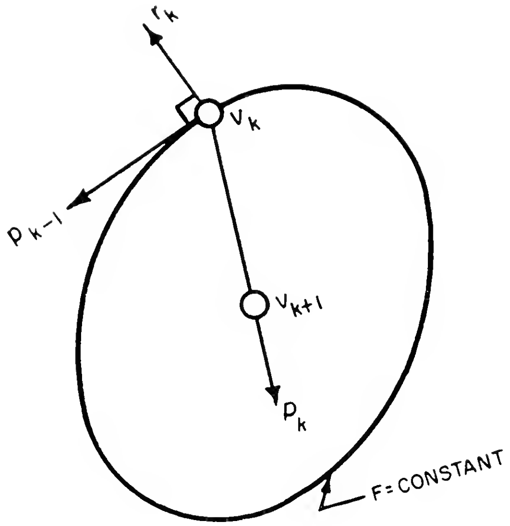


FIGURE 5.2.

Thus, in this case λ is the reciprocal Rayleigh quotient of $\boldsymbol{r}$. Occasionally, a computation is made, using formulas (11) but assuming λ to be constant instead of calculating it out at each step, the so-called weak method of steepest descent. A special case of this procedure is Liebmann's method for the numerical solution of Laplace's boundary value problem.

5.3. A still better choice of $\boldsymbol{p}$ can be made by directing it toward the center of the ellipses. Figure 5.2 shows the situation in greater detail. Let $\boldsymbol{v}_k$ be the kth trial point, which was determined with the help of the direction vector $\boldsymbol{p}_{k-1}$, and let the corresponding residue vector be $\boldsymbol{r}_k$. On passing a two-dimensional plane through $\boldsymbol{v}_k$ and $\boldsymbol{r}_k$ (the plane of fig. 5.2), the surfaces F=constant again give a set of contour lines. As we saw in the first paragraph, it is characteristic of a step in the relaxation that the ellipse passing through $\boldsymbol{v}_k$ is tangent to $\boldsymbol{p}_{k-1}$. It is now desired that the next direction vector pass through the center of the ellipse; that is, $\boldsymbol{p}_k$ and $\boldsymbol{p}_{k-1}$ must be conjugate directions in respect to the ellipse, or

$$\boldsymbol{p}_k\cdot D\boldsymbol{p}_{k-1}=0. \tag{13}$$

If we put

$$\boldsymbol{p}_k=-\boldsymbol{r}_k+\epsilon_{k-1}\boldsymbol{p}_{k-1}, \tag{14}$$

then

$$-\boldsymbol{r}_k\cdot D\boldsymbol{p}_{k-1}+\epsilon_{k-1}\boldsymbol{p}_{k-1}\cdot D\boldsymbol{p}_{k-1}=0$$

or

$$\epsilon_{k-1}=\frac{\boldsymbol{r}_k\cdot D\boldsymbol{p}_{k-1}}{\boldsymbol{p}_{k-1}\cdot D\boldsymbol{p}_{k-1}}. \tag{15}$$

From formulas (8), (9), and (10), we find for the new trial point

$$\boldsymbol{v}_{k+1}=\boldsymbol{v}_k+\lambda_k\boldsymbol{p}_k \quad \boldsymbol{r}_{k+1}=\boldsymbol{r}_k+\lambda_k D\boldsymbol{p}_k$$

with

$$\lambda_k=-\frac{\boldsymbol{r}_k\cdot\boldsymbol{p}_k}{\boldsymbol{p}_k\cdot D\boldsymbol{p}_k}. \tag{16}$$

In order to start the calculation, a first trial point $\boldsymbol{v}_0$ may be arbitrarily chosen, the corresponding residue vector computed, and the first direction vector taken as $\boldsymbol{p}_0=-\boldsymbol{r}_0$. *An interesting property of this relaxation method is that the solution of the n equations in n unknowns will be found in precisely n steps.* In other words, we have not only a method of successive approximation, but also an algebraic scheme for the exact solution.[4]

[4] Proof of this property, as well as a further simplification of the method, may be found in the article cited in footnote 3.

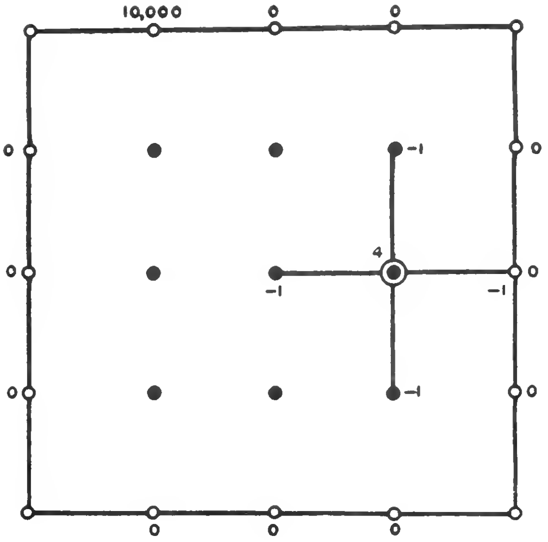


FIGURE 5.3.

TABLE 1.

k	p_k	Dp_k	r_{k+1}	p_k	Dp_k	r_{k+1}	p_k	Dp_k	r_{k+1}
			−10000			0			0
0	+10000	+40000	0	0	−10000	−2500	0		0
1	+ 1250	0	0	+2500	+ 8750	0	0	−2500	−714
2	+ 306	0	0	+ 612	− 1	0	+714	+2244	0
3	+ 93	0	0	+ 186	0	0	+217	0	0
0	+ 20	0	0	+ 40	− 1	0	+ 47	+ 1	0
	2991			*982*			*312*		
			0			0			0
0	0	−10000	−2500	0		0	0		0
1	+2500	+ 8750	0	0	−5000	−1429	0	0	0
2	+ 612	− 1	0	+1429	+4492	0	0	−2143	−682
3	+ 186	0	0	+ 434	0	0	+682	+2077	0
4	+ 40	− 1	0	+ 94	+ 2	0	+147	− 1	0
	982			*625*			*268*		
			0			0			0
0	0		0	0		0	0		0
1	0	−2500	−714	0	0	0	0	0	0
2	+714	+2244	0	0	−2143	−682	0	0	0
3	+217	0	0	+682	+2077	0	0	−1364	−448
4	+ 47	+ 1	0	+147	− 1	0	+448	+1498	0
	312			*268*			*134*		

$(r_0,r_0)=100\cdot 10^6$ $(p_0,Dp_0)=\ 400\cdot 10^6$ $\lambda_0=0.\ 25$
$(r_1,r_1)=12.5\cdot 10^6$ $(p_1,Dp_1)=4375\cdot 10^6$ $\lambda_1=0.\ 28571$ $\epsilon_0=0.\ 125$
$(r_2,r_2)=3061633$ $(p_2,Dp_2)=\ 9622276$ $\lambda_2=0.\ 31818$ $\epsilon_1=0.\ 24493$
$(r_3,r_3)=\ 930248$ $(p_3,Dp_3)=\ 2833028$ $\lambda_3=0.\ 32836$ $\epsilon_2=0.\ 30384$
$(r_4,r_4)=\ 200704$ $(p_4,Dp_4)=\ \ 671012$ $\lambda_4=0.\ 29911$ $\epsilon_3=0.\ 21575$

This scheme may also be used to invert the matrix (a_{ik}) of eq (1). The elements of the inverse matrix are

$$\delta_{ik}=\sum_{p}\frac{\boldsymbol{p}_{pi}\boldsymbol{p}_{pk}}{\boldsymbol{p}_{p}\cdot D\boldsymbol{p}_{p}},$$

where p_{pi} is the ith component of the pth direction vector. Furthermore, the successive residue vectors $\boldsymbol{r}_0$, $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, . . . form an orthogonal system, a property that may be used to calculate the characteristic values of the matrix (a_{ik}). The resulting procedure is similar to that suggested by Lanczos.[5]

Table 1 contains a simple example, a numerical solution of Laplace's equation with the boundary values shown in figure 5.3. The function $\boldsymbol{u}$ is to be determined for the nine inner points, such that $D\boldsymbol{u}=0$, where D is the operator given by the small cross $(4, -1, -1, -1, -1)$. Each box in the table corresponds to one of the nine points and contains the successive values of $\boldsymbol{p}_k$, $D\boldsymbol{p}_k$, and $\boldsymbol{r}_{k+1}$. The procedure was started with $\boldsymbol{v}_0\equiv 0$. The values of the desired function $\boldsymbol{u}$ are given at the end.

This method is especially suitable when D is a difference operator, such as results from the approximation of a partial differential equation by a difference equation. The calculation of $D\boldsymbol{v}$ is then not so long as it would be if D were a more general matrix (a_{ik}). At the present time, this method is being used on the Zurich Relay Calculator to compute the elastic deformations of a plane plate under 29 different loading conditions. The results of each step in the relaxation are punched into tape and fed back into the machine again for the next step.

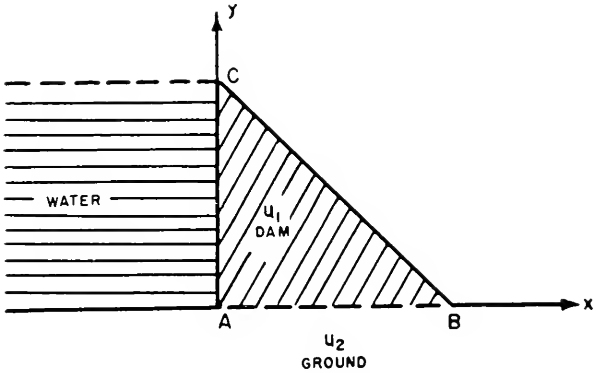


FIGURE 5.4.

5.4. In a somewhat different manner the following problem has been handled on the Zurich machine. The Airy stress function is to be calculated for the profile of a dam supported elastically by the ground (fig. 5.4). The ground is assumed to be a half-plane. The problem yields the partial differential equation $\Delta\Delta u=0$ and the following boundary conditions between the stress functions u_1 in the dam and u_2 in the ground:

$$u_1=u_2, \qquad \frac{\partial u_1}{\partial y}=\frac{\partial u_2}{\partial y}$$

$$\frac{\partial^2 u_2}{\partial y^2}=\frac{4}{15}\frac{\partial^2 u_1}{\partial y^2}+\frac{13}{45}\frac{\partial^2 u_1}{\partial x^2}$$

$$\frac{\partial^3 u_2}{\partial y^3}=\frac{4}{15}\frac{\partial^3 u_1}{\partial y^3}-\frac{79}{45}\frac{\partial^3 u_1}{\partial x^2\partial y}+\frac{13}{18}.$$

To solve the problem, the triangle ABC was covered by a grid containing 139 points. The difference operator relating the value of the function at each grid point to the values at neighboring points is shown in figure 5.5. In the ground below the dam, the explicit solution developed by Boussinesq for the boundary value problem $\Delta\Delta u=0$ in a half-plane was used, which required some modification of the

[5] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Research NBS 45, 255–282 (1950) RP2133.

difference operator along the line AB. The 139 equations in 139 unknowns were solved by using a relaxation method related to the one in paragraph 5.3, in which an approximate computation of the first eigenfunction was used to eliminate the contribution of that function to the residual errors. . The work was carried out on the machine in blocks of 16 grid points, whose residues were simultaneously reduced to zero. Incidentally, by an application of the theory of group characters, these 16 equations could be reduced to two sets of three equations each plus four other equations. Each such block took 30 minutes of machine time and the entire calculation to five significant figures about 100 hours.

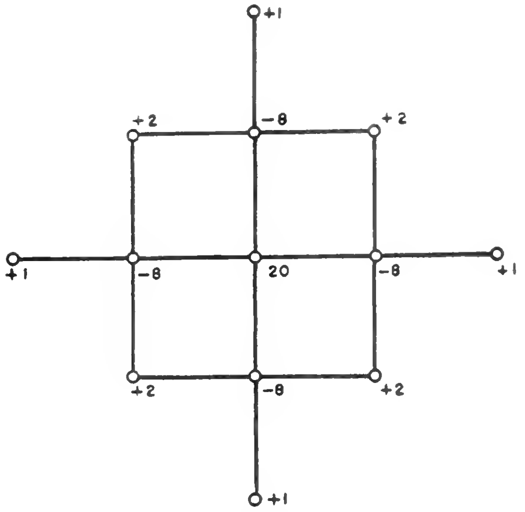


FIGURE 5.5.

6. Errors of Matrix Computations

Paul S. Dwyer [1]

Introduction

Considerable work has been done during recent years on the general question of the accumulation of errors when sets of approximate simultaneous linear equations are solved by approximate methods. I refer particularly to the work of Etherington [1],[2] Tuckerman [2], Lonseth [3, 4], Hotelling [5, 6], Satterthwaite [7], von Neumann and Goldstine [8, 9], Turing [10], and Ostrowski [11], as well as to the work of others mentioned in the list of references. My list of references includes about 30 titles and is not as inclusive as is the list of G. E. Forsythe [12], though the area covered in my list corresponds in a general way to that of his last main subdivision IV.

It is my purpose here to present a discussion of the nature of the general mathematical problem referred to above, together with indications as to the various methods proposed for handling the different phases of it. Perhaps I should add that this paper is related to the material in chapters 2 and 17 of my book on Linear Computations [13], although the presentation here is more general since it is not directed specifically toward problems involving a relatively small number of variables. Problems involving few variables can be handled with desk machines by exact pivotal methods or by approximate pivotal methods which do not involve extensive accumulations of rounding off errors.

The problem as proposed leads to an examination of the nature of the fundamental errors involved. This in turn demands an examination of the numbers used to express these errors. The question of how these fundamental numbers can be combined leads us, in the general matrix statement of the problem, to an examination of the errors of addition, subtraction, and multiplication, as well as inversion. It seems appropriate then to use the title, "Errors of Matrix Computations", even though the major effort of the paper is devoted to the question of the errors of the solutions of simultaneous linear equations, and the closely related problem of the errors of the calculated inverse matrix.

Types of Errors

Von Neumann and Goldstine [8: 1023–1032] have discussed the errors involved when a problem in pure or applied mathematics is solved by numerical computation. These errors are errors due to theory, errors due to observation, errors of approximation or truncation, as well as rounding-off errors. It is not my purpose to discuss all of these errors. I direct your attention to the situation in which the theoretical statement of the problem is accepted, but the mathematical formulation of the problem involves coefficients which are subject to unspecified but limited error. For example, we wish to discuss the errors of the solution of a set of simultaneous linear equations whose coefficients are subject to error. We do not question the suitability of the linear statement of the problem, but we are concerned (1) with the effect of the original errors which are inherent in the statement of the problem and (2) with the cumulation of the errors which result from the continued use of approximate methods, if approximate methods are necessary. We will call these latter errors the *computational errors*, and we will follow Milne [19: 30] in calling the original errors the *inherent errors*.

As an illustration we consider the matrix form of simultaneous linear equations with coefficients subject to error as

$$[a+\epsilon(a)]x=f+\epsilon(f), \tag{1}$$

[1] Professor of Mathematics, University of Michigan.

[2] Figures In brackets indicate the literature references on p. 58, and numbers following the colon indicate page number.

where x and f are column vectors, and where $\epsilon(a)$ and $\epsilon(f)$ are the inherent errors of the elements of the matrices a and f, respectively. If $\epsilon(a)$ and $\epsilon(f)$ are known, there is in general a unique solution to (1). However, if approximate methods are used in the solution, computational errors will be introduced so that the proposed solution for x will not satisfy (1) exactly. The cumulation of these computational errors has been a chief concern of several recent writers.

In the more typical problem the values of $\epsilon(a)$ and $\epsilon(f)$ are not known but are limited by some bounds (which are usually small compared with the elements to which they correspond). In this case it is essential that we learn what effect these unspecified but limited errors have on the values of the resultant solution.

We thus have two problems confronting us: (1) What can we say about the effect of the inherent errors on the solution? (2) What can we say about the effect of the cumulations of the computational errors on the solution if approximate methods are used?

In many situations the inherent errors as well as the computational errors result from rounding off so that these two types of errors are treated simultaneously as rounding-off errors. It is a chief contention of this paper that this should not be done. The two types of errors should be treated separately insofar as that is possible, and the results may then be integrated to provide a suitable answer to the mathematical problem.

Use of Significant Digits

We limit our discussion to finite real numbers. We further limit our discussion to digital numbers. It is not true that every real number can be written in pure digital form, but every real number can be put in the form

$$N=\bar{N}+\epsilon, \tag{2}$$

where $\bar{N}$ is a digital number. Furthermore, ϵ can be replaced by a digital positive number η, whose value is equal to or greater than the absolute value of ϵ. Thus every finite real number can be approximated by a digital number with its digital bound. The digital number $\bar{N}$ and the digital bound η locate the real number on the interval $\bar{N}-\eta$ to $\bar{N}+\eta$. Hence any nondigital real number can, with some rounding off, be written in the same form as an approximate number, that is, one whose value is known to be in the scale interval bounded by $\bar{N}-\eta$ and $\bar{N}+\eta$.

There is nothing new about this, of course, for it is the very basis of the conventional use of significant figures. When using significant figures the convention adopted is that the η is one-half of the last recorded decimal unit, so that the value of $\bar{N}$ appears but not the bound. This leads to a simple computing technique using digital numbers, but it is a very unsatisfactory technique because the convention limits many of the approximate numbers which should be used in expressing precise though approximate results to a small subclass of all approximate numbers. No approximate number which has a bound greater than one-half in the last recorded decimal position is permitted.

An approximate number has two components, and any trick that is used to restrict the number to one component by holding the other component constant is very unsatisfactory. The conventional rules for working with significant digits [13:27–33, 14:2–24] are so unsatisfactory that they provide no guarantee that the results of a continued series of arithmetic operations, such as one finds in the problem of the solution of simultaneous equations, will have any meaning at all.

Further study with more adequate systems of approximate numbers shows that the unsatisfactory results of computations using numbers with significant digits are usually due more to the deficiencies of the inadequate tools we have selected to work with than they are to the inherent indeterminacy of the problem.

More adequate approximate numbers are necessary and these will be discussed shortly. First, there are one or two things which I think we might learn from our experience with computation with significant figures.

Significant figures were designed to handle approximate numbers in an easy way. Perhaps this was in order before we had machines to do our computing, but certainly it is not in order now. There is no point in discarding vital information just to simplify the computation, but that is what we are doing continually when we use the general laws of computing with significant digits.

Computation with significant digits features over-all rules by which the number of significant places can be obtained with a trivial amount of computational work. Such general rules are not of much value in connection with a specific problem. Considerable additional computational work is demanded to compute the two components of each approximate number needed for the solution of the problem, but usually the results are very much more precise than are the answers obtained by the rules of computation with significant digits.

We have given up the formal rules of computation with significant digits long ago, but I feel we are still trying to get over-all rules that will cover the accumulation of rounding-off errors. These over-all rules are worth while in understanding the nature of the processes involved but, in general, they should not be expected to give results which are precise enough for a given individual problem. If we have a problem with coefficients subject to error, let us study the accumulation of the errors in the problem itself. Of course, this may take considerable extra computation, but that is the way to get the precise answer to our problem, if we want it! Besides, we are now getting simple and complex machines to do this computing for us.

I propose then that we should expect that specific answers to a problem that is expressed in terms of numbers whose coefficients are subject to error should be answered in terms of the specific elements of the problem. The practical question is not how large may the accumulation of the errors be in problems of a given type but how large is the accumulation of errors in this specific problem!

In the main, the formulas in the later sections of this paper feature the specific problem rather than the general result. If we have to compute a certain inverse matrix, for example, in order to answer certain questions, let's compute it!

Types of Approximate Numbers

I have indicated above that the treatment of approximate numbers based on significant figures is very unsatisfactory. Any satisfactory representation of an approximate number must permit the precise identification (within digital limitation) of the complete approximate number as an interval on a scale. This can be done by giving the highest and lowest points of the interval, these being the two components of what I have called a *range number* [13:12], or it can be done by specifying the midpoint of the interval, which we might call the point approximation, together with the maximum error possible from this point approximation. This number I have called an *approximation-error number* [13:12]. It is easy, of course, to change from range numbers to approximation-error numbers, and vice versa.

Modifications in these basic concepts are possible. The pure range numbers may be replaced by fiducial range numbers where the nature of the problem permits the estimation by probability theory. This concept may be extended to include statistical approximate numbers where the approximate number is indicated by a mean and a variance with perhaps some distribution law for the unspecified error.

The approximation-error numbers may be replaced by first-order approximation-error numbers when first-order error terms only are used in their calculation.

In many applied problems the errors are not known, but bounds for the errors are known. The errors of approximation-error numbers are usually given in terms of these bounds. It is proper to use rounding off in expressing these bounds so that the error component, as well as the point approximation, is expressed in terms of digital numbers. These dual components can be rounded simultaneously to a fixed number of places. Thus the approximation-error number 1.9684(186) can be rounded off successively to get 1.968(19) and 1.97(2).

This double rounding-off method is satisfactory in replacing an approximate number by another less precise approximate number which does not differ from the first except in units of the discarded digits. This is also satisfactory if first order approximation-error numbers are used. On the other hand, if one wishes to guarantee that the range of the rounded-off approximate number includes all of the range of the original number, he may express the original number in range form and then round off, not necessarily to the nearest digit, but so as to include the original range. Thus the approximation-error number above

$$1.9684(186) = \begin{bmatrix} 1.9870 \\ 1.9498 \end{bmatrix}$$

rounds off successively to $\begin{bmatrix} 1.987 \\ 1.949 \end{bmatrix}$ and thence to $\begin{bmatrix} 1.99 \\ 1.94 \end{bmatrix}$.

In some cases the point approximation of the approximation-error number alone may be used. I have called numbers of this type *incomplete approximation-error numbers* [13:34], or *incomplete numbers* for short, to point out the fact that results obtained by using these numbers are not complete until bounds for the errors are added. These numbers can be carried to a fixed number of places, as is desirable in machine calculation, without implication that the figures recorded are significant in the technical sense. Some computers call the figures of these numbers "significant figures" and reserve the term "determinate significant figures" for those that are significant in the technical sense [2]. The incomplete numbers are exact digital numbers.

Approximate Matrices

A matrix whose elements are approximate numbers might be called an *approximate matrix*. The elements of these matrices might appear either in range or approximation-error form. In general, such an m by n approximate matrix is composed of $2mn$ element components, since each of the elements in an exact matrix is replaced by dual components in the approximate matrix. Approximate matrices may be rounded off by rounding off the dual components of each element according to the methods described in the last section.

A matrix whose elements are incomplete numbers, that is, a matrix in which the digital point approximation components alone appear, might be called an *incomplete matrix*. The complete approximate matrix calls, in addition, for a bound for the error component for each term of the matrix.

Operations With Approximate Matrices

Rules for computation with approximate numbers in range and approximation-error form are available [13: 16–25], although recommended rules for calculating the errors of approximation-error numbers are really the rules of differential calculus and hence lead to first-order approximation-error numbers. These rules can be applied, of course, to operations with matrices whose elements are approximate numbers. With careful attention to details it is possible to establish suitable rules for sums, differences, and products of approximate matrices when the matrices are expressed either in range or approximation-error form. The algebraic operation of division leads to the matrix operation of pre- or post-multiplication by the inverse. The problem of inverting an approximate matrix is closely related to that of the solution of approximate equations, which is discussed in some detail in the later sections of this paper.

In general, operations with approximate matrices give results which are approximate matrices with guaranteed bounds. This is certainly true if precise methods are used in getting the bounds at each stage of the computation. It is approximately true when first-order approximation-error methods are used in the computation if the relative error of each calculated term is small. These bounds are not guaranteed, at least in the sense used above, when some assumption is made about the distribution of the true error in the approximation-error numbers and statistical formulas are used to estimate the range at some probability level.

The strict results are reasonably satisfactory when the problem demands continued additions and subtractions, although the probability approach, under suitable assumptions, may lead to closer practical bounds [15], even though these bounds can not be shown to be true in the strict sense.

The strict results are not so satisfactory for operations involving continued multiplications though, theoretically, all multiplications can be carried out with the use of range numbers or approximation-error numbers, and the commutative property of product operations is preserved. However, an extensive series of multiplications leads to numbers with so many places that it is necessary, from a practical standpoint, to round off. It is desirable, in such cases, to round off to a fixed digital position. The round-off procedure introduces rounding-off errors that differ according to the particular product involved, so the final product, when the intermediate products have been rounded to some given decimal position, varies with the order in which the different factors are included in the accumulating product. Multiplication by approximate numbers in appropriate form is commutative, but it is not commutative if, for the sake of simplification, we indulge in rounding off in the process of computation. In a loose sense, we might say that commutativity is approached as the rounding-off errors are made smaller and smaller.

First-order multiplications with approximation-error numbers are not in general commutative since second-order terms are neglected at each multiplication. However, if the relative errors are small, the second-order relative errors are very small and the differences between the various products are trivial.

Formal treatment of material very similar to this is given by von Neumann and Goldstine [8 : 1035–1039], in their discussion which shows that in general pseudo-operations are not commutative.

The situation is somewhat worse with respect to division because the quotient of two exact digital numbers is not necessarily digital. It follows that the point approximation of the ratio of two approximate numbers is not necessarily digital but must itself be rounded off to attain digital form. It is obvious that the final result of a series of divisions depends upon the order in which the divisions are performed if the divisions are carried to a fixed decimal position.

Difficulties of a much larger order are encountered when division by an approximate number is replaced by the inversion of an approximate matrix. I would like to devote the most of the remainder of my space to this problem and to the allied problem of the numerical solution of equations with coefficients subject to unspecified but limited error.

Before taking up this topic, I would like to make one remark about the cumulation of inherent error in calculation. Calculations with range numbers or with approximation-error numbers are very cumbersome when numerous successive calculations are demanded. It is preferable to calculate incomplete numbers and, after the answer is given in terms of incomplete numbers, to calculate bounds for the answer with the use of an appropriate error formula, if such a formula is available. Such a formula can sometimes be derived from the theoretical statement of the problem, if care has been taken in stating the problem so that the inherent errors are separated from the calculational errors. This general method is illustrated below by applying it, in considerable detail, to the linear equations problem.

The Solution of Approximate Simultaneous Equations

Just what do we mean by the solution of linear simultaneous equations with coefficients subject to unspecified but limited error? First, what do the equations themselves mean? These equations do not represent just one set of equations, but a multiple infinity of sets of equations, since each coefficient that is indicated by an approximate number in digital form actually might be any one of an infinity of values. Thus the algebraic equations

$$(a_{11}+\epsilon_{11})x_1+(a_{12}+\epsilon_{12})x_2=a_{13}+\epsilon_{13}$$

$$(a_{21}+\epsilon_{21})x_1+(a_{22}+\epsilon_{22})x_2=a_{23}+\epsilon_{23}$$

represent a sextuple infinity of sets of two equations in two unknowns, since each ϵ_{ij} may be taken on any of the infinity of values between $-\eta_{ij}$ and η_{ij}.

As the multiple infinity of the sets of equations can only be represented symbolically, it is obviously impossible to write all the multiple infinity of solutions explicitly. What we can do is provide some single solution that may be said to represent the multiple infinity of all solutions. This may be done in various ways.

We may obtain the unique solution of any one of the multiple infinity of sets of exact equations and use that solution as representative of all the solutions. This may be satisfactory in some cases, but it must be noted that the solution is not unique and that the answer provides no measure of the extent to which this solution may differ from the solution of some other set. The particular set of equations chosen may have a solution that is extreme, so that the solution is not very satisfactory as a representative one.

An arbitrary unique solution that may have some special merit is that obtained by solving the set of equations in which all the error terms are zero. The resulting equations are unique exact equations, and they may be solved to any desired degree of accuracy. The results are incomplete numbers, but these may be said, in a unique sense, to represent all the multiple infinity of solutions. This type of solution, unless modified, is also quite unsatisfactory because there is no indication of the extent to which the other solutions may differ from this one. We might call this the *incomplete solution*, since it is the unique solution of the incomplete equations.

A more satisfactory type of solution is that in which the answers are given as approximate numbers. Not only is the point approximation given, but some estimate of the variation of the answer is provided.

With the use of suitable formulas derived separately, and with additional computation, we can provide bounds for the answer to the incomplete solution just described. This can be done approximately with the use of first-order error terms or, more precisely, with higher order error terms. The extrema of the various x_i can be computed or, if desired, the values of the other x_j associated in a solution with each extreme x_i can be computed.

Under suitable assumptions, probability formulas can be used to indicate the expected deviation of a typical solution from the incomplete solution.

In consideration of the above discussion it seems desirable to express the solution, as well as the coefficients, in terms of approximate numbers. The mathematical formulation of (1) might then be modified to allow for the error in the solution and would appear as

$$[a+\epsilon(a)][x+\epsilon(x)]=f+\epsilon(f), \tag{3}$$

where x indicates the incomplete solution, which is the solution of the incomplete equation

$$ax=f, \tag{4}$$

if a is nonsingular, and can be indicated by

$$x=a^{-1}f. \tag{5}$$

Since the elements of a and f are exact digital numbers, it follows that the value of each x_i can be written as the ratio of two digital numbers, as is evident from an application of Cramer's Rule.

Simultaneous solution of the matrix eq (3) and (4) enables us to get a theoretical expression for $\epsilon(x)$. We get

$$a\epsilon(x)+\epsilon(a)x+\epsilon(a)\epsilon(x)=\epsilon(f) \tag{6}$$

so that

$$[a+\epsilon(a)]\epsilon(x)=\epsilon(f)-\epsilon(a)x \tag{7}$$

$$\epsilon(x)=[a+\epsilon(a)]^{-1}[\epsilon(f)-\epsilon(a)x]. \tag{8}$$

A series expansion of (8) results in the formula:

$$\epsilon(x)=\{I-a^{-1}\epsilon(a)+[a^{-1}\epsilon(a)]^2-[a^{-1}\epsilon(a)]^3+. . .\}a^{-1}[\epsilon(f)-\epsilon(a)x],^3 \tag{9}$$

which is similar to matrix formulas used by Lonseth [3], Turing [10], and Waugh [16].

The formula provides the theoretical answer for the errors of the solution if the original inherent errors are known and are relatively small. It can be used as the basis of a formula for the bound of the errors of the solutions when, as is usual, the inherent errors are indicated by their bounds. Various formulas for bounds are discussed later. First, we examine the question of the solution of the incomplete equations.

[3] Valid under favorable conditions of convergence.

The Solution of the Incomplete Equations

The theoretical solution of the incomplete equation (4) is very simple, if a is nonsingular, but the numerical solution becomes complicated if many variables are involved. The numerical solution of these equations seems to have received more attention during recent years than has any other phase of the general problem. A chief interest has been the accumulation of the rounding-off errors with approximate methods [5] and the fact that in general no guarantee can be given that these cumulations are not enormous [5]. I should also mention studies of particular methods with a view to selection of a method in which the cumulation of rounding-off errors will be a minimum [9, 10], the use of iteration as a means of improving the accuracy which may be destroyed by successive round offs [5, 6], the use of error control in which matrix manipulations are applied to the original matrix equation so as to reduce it to a form in which the round-off accumulations are known to be relatively trivial [7], the use of functions of the elements of a matrix such as the norm in obtaining bounds for the accumulation of the rounding-off error [5, 6, 7, 8, 10, 11], the use of scale adjustments to bring the elements of the matrix into comparable ranges [8: 1034; 17: 30; 18], etc.

I would like to discuss these in detail, but space permits only a passing reference to these important results and techniques.

My experience has led me to the conclusion, and there are some others who have arrived at the same conclusion [10], that direct elimination methods are for many problems best. Certain types of problems converge quickly with iterative methods, and we do not know yet just what methods we will use when we solve equations in very many variables with various electronic digital computers but, with present problems and equipment, direct-pivoted methods seem not too unsatisfactory if we can handle the problem of the accumulation of errors.

Where the number of variables is relatively small, I like to use exact pivotal elimination methods [13: 50–89]. There is some theoretical advantage in using a modified form of an exact method when the number of variables is fairly large [13: 88], since each calculated value can be interpreted as a determinant.

It is perhaps preferable to select the pivots from the diagonal terms since they are the nonzero multipliers of principal minors, and it is helpful to know that these principal minors do not vanish.

For present purposes, I assume that some approximate solution to (4) can be obtained with the use of some one of the various methods described by Forsythe [12]. It is desired to know how satisfactory this approximate solution is. It is not necessary that the solution be exact, but it is important that the errors of the solution of the incomplete equation are not larger than the accumulation of the inherent errors.

We can substitute the proposed solution in the left hand side of the incomplete equations to see how well the equations are satisfied. If we let e indicate the matrix of computational errors, we can write

$$a[x+e(x)]=f+e(f). \tag{10}$$

If the values of $e(f)$ are within the bounds specified for $\epsilon(f)$ we may say in general that the solution is satisfactory, unless the equations are *ill-conditioned*. Methods of studying the extent of ill-conditioning of a matrix have been developed, see, for example [10], but it is my thesis that ill-conditioning will in general be revealed if direct elimination methods are used with diagonal pivots and certainly if the inverse matrix is calculated. In fact, formulas for ill-conditioning used by Turing [10] feature the elements of the inverse.

If the values of e are too large, it is possible to follow the iterative methods of Hotelling [6], the error control method of Satterthwaite [7], or, to use the approximate inverse, c_0, of a as the basis of a further approximation. Substitution of the fact that $ax=f$ in (10) yields $ae(x)=e(f)$, so that

$$e(x)=a^{-1}e(f) \tag{11}$$

with $a^{-1}=c_0+e(c_0)$. We have $e(x)=c_0e(f)+e(c_0)e(f)$ and

$$e(x)=c_0e(f) \tag{12}$$

as the first-order approximation. Other methods could be used, but this uses the approximate inverse.

It should be noted that all of these methods, including the examination of ill-conditioning, call for the calculation of the inverse matrix or an approximation to it.

Bounds for the Inherent Errors of the Solutions

Bounds can be found for the errors indicated by (9). If we omit the error terms of order higher than one, we have

$$\epsilon(x)=a^{-1}[\epsilon(f)-\epsilon(a)x]. \tag{13}$$

We get values for the bounds of $\epsilon(x_k)$ when the absolute values of all the terms in (13) are taken and the $\epsilon(a)$ and $\epsilon(f)$ are replaced by their bounds. For example, if η is the bound for all the elements of $\epsilon(a)$ and $\epsilon(f)$, we obtain the formula

$$\eta\,(x_k)=\eta\,(1+\sum|x_i|)\sum_j|a_{ki}^{-1}|, \tag{14}$$

where the pairs of vertical lines indicate absolute values. This formula, or formulas very similar to it, have been given by several authors [1, 3, 13:262].

It should be pointed out that the first-order answer to the problem is now stated in terms of the values of x_i, which are the solutions of the (exact) incomplete equations, and the elements of a^{-1}, which is the inverse of an exact matrix. This is very desirable because the formula gives bounds for the effect of the inherent errors which do not involve the computational errors. To be sure, we use values which are approximations to those of x_i and a^{-1}, but, in practice, these approximations need be carried only far enough to fix the first decimal positions of $\eta(x_k)$.

The formula (14) does give first-order bounds for the extreme errors of the different x_k, but, in a real sense, it does not give solutions since $\eta(x_k)$ is not necessarily associated with $\eta(x_j)$. It is possible to return to (13) and to assign values to the $\epsilon(f)$ and the $\epsilon(a)$ in such a way as to maximize some particular $\eta(x_k)$, and, by substitution in (1) to produce the actual equations, aside from errors of order higher than the first, which have these extreme solutions. Techniques that use the inverse or adjugate matrix are available for determining these sets of extreme equations [13: 264–276].

The formulas above use the inverse matrix. Some authors prefer to find bounds for first-order errors without the calculation of the inverse matrix. Among these are Milne [19:29–35] and Willers [20:271–273], who use, in effect, the formula

$$a\epsilon(x)=\epsilon(f)-\epsilon(a)x, \tag{15}$$

which is (13) premultiplied by a. The right-hand side is replaced by its bound $(1+\Sigma|x|)\eta$. Willers and Milne use different methods of solution in guaranteeing the bounds and in general arrive at bounds that are much larger than they need be. These bounds become very much larger than the least upper bound if a back solution is used. It is wise, if close bounds are desired, to carry through a new elimination process for each x_k, and, if this is done, one might as well calculate the inverse matrix and use the adequate bounds that this method provides.

The formula (14) simplifies somewhat when all the $\epsilon(f)$, all the $\epsilon(a)$, or certain of the values of $\epsilon(a)$, are zero [13:278–284].

Higher order approximations to the errors can be obtained with the use of more of the terms of (9). Thus the second approximation is

$$\begin{aligned}\epsilon(x)&=[I-a^{-1}\epsilon(a)]a^{-1}[\epsilon(f)-\epsilon(a)x]\\&=a^{-1}[\epsilon(f)-\epsilon(a)x]-a^{-1}\epsilon(a)a^{-1}[\epsilon(f)-\epsilon(a)x].\end{aligned} \tag{16}$$

Bounds can then be taken as before.

Bounds for second-order errors have been provided by Lonseth [3], who has also used a recursion formula and mathematical induction in estimating bounds for higher order errors. Lonseth [3] and Etherington [1] have made some study of suitable probability distributions of these errors.

The Errors of the Inverse

The discussion of the errors of the inverse is rather brief, because the calculation of the inverse is, essentially, a special case of the more general problem of the solution of sets of simultaneous equations. It results from solving the equation

$$[a+\epsilon(a)][x+\epsilon(x)]=I. \tag{17}$$

The identity matrix is not subject to error so the problem is somewhat less complicated than the previous one. We obtain the incomplete solution from the exact incomplete equations

$$ax=I \tag{18}$$

as before and have

$$\epsilon(a)x+a\epsilon(x)+\epsilon(a)\epsilon(x)=0. \tag{19}$$

It follows that

$$\epsilon(x)=-[a+\epsilon(a)]^{-1}\epsilon(a)x=-[a+\epsilon(a)]^{-1}\epsilon(a)a^{-1}. \tag{20}$$

This is expanded to get

$$\epsilon(x)=-[I+a^{-1}\epsilon(a)]^{-1}a^{-1}\epsilon(a)a^{-1}. \tag{21}$$

The first-order approximation is

$$\epsilon(a^{-1})=-a^{-1}\epsilon(a)a^{-1}. \tag{22}$$

The formula, which is a matrix generalization of the formula for the differential of x^{-1}, expresses the inherent errors of the inverse in terms of the original inherent errors and the elements of the incomplete inverse, which appears twice in the numerator.

Bounds can then be determined for each of the elements of the inverse from formulas that feature the absolute values of terms of the incomplete inverse [13:284–288].

Summary and Conclusions

This paper results from an attempt to study the nature of the general mathematical problem of the solution of approximate simultaneous equations and to indicate theoretical and practical solutions by analyzing the problem in such a way that the effects of the inherent errors are separated from the computational errors. The presentation features references to much important work that has been done on several phases of the problem. Much of this basic work, which is barely mentioned, deserves more detailed consideration than the time limits of this paper have permitted.

A general discussion of the nature of the errors involved is followed by an examination of suitable means of expressing the results. A matrix presentation of the problem is useful, if the elements of the matrix are range numbers or approximation-error numbers. The single-component numbers that feature significant digits are not satisfactory.

Extensive computations with range numbers, or with approximation-error numbers, are unsatisfactory, in practice, when long series of operations are demanded, as in the solution of linear equations with pivotal methods. The necessity of rounding off, and the alternative selection of pivots, leads one to ranges and bounds that are not as close as is possible, although they include the answers.

These difficulties are largely avoided with the use of incomplete numbers when a separate formula for determining the error bounds is available. The matrix statement of the problem of linear equations is broken down into an exact matrix equation whose unknowns are x_i, the values of the incomplete solution, and another matrix equation whose unknowns are the $\epsilon(x_i)$, the values of the inherent errors of the solution. The formula for $\epsilon(x_i)$ features the inverse of the matrix of the coefficients, a^{-1}, of the incomplete equations, as well as the inherent errors, ϵ_{ij}, of the coefficients.

Any approximate solution of the incomplete equations will result in values $x_i+e(x_i)$, where x_i is the exact value, and $e(x_i)$ is the accumulation of the rounding-off errors. From the practical standpoint, a reasonably satisfactory solution is obtained when $e(x_i)$ is appreciably less than $\eta(x_i)$, the value of the bound for $\epsilon(x_i)$. Less precisely, an approximation may be considered to be satisfactory when $e(f)$ is very small. A more satisfactory first-order approximation for $e(x)$ may be obtained by premultiplying $e(f)$ by c_0 an approximation to a^{-1}.

The formulas of this paper are specific formulas in that they feature the elements of a^{-1}. They demand the calculation of a^{-1} or of some good approximation to it. Rough approximations are adequate for certain practical problems, and over-all rules are then in order, but we should clearly distinguish such over-all rules from mathematically precise solutions.

I hope that these remarks will help to clarify some of the various questions that arise in the study of simultaneous linear equations with unspecified but limited error.

References

[1] I. M. H. Etherington, On errors in determinants, Proc. Edinburgh Math. Soc. [2] **3,** 107–117 (1932).

[2] L. B. Tuckerman, On the mathematically significant figures in the solution of simultaneous linear equations, Ann. Math. Stat. **12,** 307–316 (1942).

[3] A. T. Lonseth, Systems of linear equations with coefficients subject to error, Ann. Math. Stat. **13,** 332–337 (1942).

[4] A. T. Lonseth, On relative errors in systems of linear equations, Ann. Math. Stat. **15,** 323–325 (1944).

[5] H. Hotelling, Some new methods in matrix calculations, Ann. Math. Stat. **14,** 1–34 (1943).

[6] H. Hotelling, Practical problems of matrix calculation, Proc. Berkeley Symposium on Mathematical Statistics and Probability, pp. 275–293 (University of California Press, 1949).

[7] F. E. Satterthwaite, Error control in matrix calculation, Ann. Math. Stat. **15,** 373–387 (1944).

[8] J. von Neumann and H. H. Goldstine, Numerical inverting of matrices of high order, Bul. Am. Math. Soc. **53,** 1021–1099 (1947).

[9] J. von Neumann and H. H. Goldstine, Numerical inverting of matrices of high order, II, Proc. Am. Math. Soc. **2,** 188–202 (1951).

[10] A. M. Turing, Rounding off errors in matrix processes, Quart. J. Mech. Applied Math. **1,** 287–308 (1948).

[11] A. Ostrowski, Simultaneous systems of equations. See paper No. 2, p. 29.

[12] G. E. Forsythe, Tentative classification of methods and bibliography on solving systems of linear equations. See paper No. 1, p. 1.

[13] P. S. Dwyer, Linear computations (John Wiley & Sons, New York, 1951).

[14] J. B. Scarborough, Numerical mathematical analysis (Johns Hopkins Press, Baltimore, 2d ed., 1950).

[15] W. S. Loud, The probability of a correct result with a certain rounding-off procedure, Proc. Am. Math. Soc. **2,** 440–446 (1951).

[16] F. V. Waugh, Inversion of the Leontief matrix by power series, Econometrika **18,** 142–154 (1950).

[17] D. B. Duncan and J. F. Kenney, On the solution of normal equations and related topics (Edwards Brothers, Ann Arbor, 1946).

[18] D. H. Leavens, Accuracy in the Doolittle solution, Econometrika **15,** 45–50 (1947).

[19] W. E. Milne, Numerical calculus (Princeton University Press, 1949).

[20] F. A. Willers, Practical analysis, translated by Robert Beyer (Dover Publications, New York, 1948).

ADDITIONAL REFERENCES

L. Fox, H. D. Huskey, and J. H. Wilkinson, Notes on the solution of algebraic linear simultaneous equations, Quart. J. Mech. Applied Math. **1,** 149–173 (1948).

G. E. Forsythe and R. A. Leibler, Matrix inversion by the Monte Carlo method, Mathematical Tables and Other Aids to Computation **4,** 127–129 (1950).

R. A. Frazer, W. J. Duncan, and A. R. Collar, Elementary matrices (Cambridge University Press, Cambridge, 1947).

P. Hartman and A. Wintner, On the effect of decimal corrections on errors of observations, Ann. Math. Stat. **19,** 389–393 (1948).

P. G. Hoel, The errors involved in evaluating correlation determinants, Ann. Math. Stat. **11,** 58–65 (1940).

F. R. Moulton, On the solutions of linear equations having small determinants, Am. Math. Monthly **20,** 242–249 (1913).

R. Redheffer, Errors in simultaneous linear equations, Quart. Applied Math. **6,** 342–343 (1948).

F. M. Verzuh, The solution of simultaneous linear equations with the aid of the 602 calculating punch, Mathematical Tables and Other Aids to Computation **3,** 453–462 (1949).

H. Walker and V. Sanford, Computation with approximate numbers, Ann. Math. Stat. **5,** 1–12 (1934).

A. W. Wundheiler, The necessity of error analysis in numerical computation, Ann. Computation Lab. Harvard Univ. **16,** 83–90 (1948).

7. Rapidly Converging Iterative Methods for Solving Linear Equations

J. Barkley Rosser[1]

1. In the fall of 1949 the Institute of Numerical Analysis of the National Bureau of Standards undertook an extended study of methods of solving simultaneous linear equations. Through the use of colloquia and discussion groups, nearly all scientific members of the Institute have made some sort of contribution to the problem. Accordingly, it is impossible to assign complete credit for the results disclosed herein to a single person or a few persons. However, certain members of the staff have given concentrated attention to the problem over an extended period and are primarily responsible for the results noted herein. In alphabetical order, these are G. E. Forsythe, M. R. Hestenes, C. Lanczos, T. Motzkin, L. J. Paige, and J. B. Rosser. Of these, the last is serving as expositor in the present paper.

2. *Finding the center of an ellipsoid.* Suppose we have an ellipsoid in n dimensions. Consider the family of ellipsoids, concentric with, similar to, and similarly oriented with the given ellipsoid. Any one of these can be considered as generated from the given ellipsoid by expansion (or contraction) from the center in a given ratio.

We first outline a theoretical scheme for finding the common center of these ellipsoids in n steps. (We give the computational details later.) We note that for any direction z_i there is a conjugate hyperplane P_i defined as the hyperplane conjugate to the point at infinity along the direction z_i. Given z_i, the plane P_i is the same for each ellipsoid of our family, and for each ellipsoid of the family the hyperplane P_i bisects all chords having the direction z_i. In particular, P_i passes through the center of all the ellipsoids.

Suppose we determine a series of directions $z_1, \ldots, z_n$, and their conjugate hyperplanes $P_1, \ldots, P_n$, with the property that z_i is parallel to each of $P_1, \ldots, P_{i-1}$. Then we can determine the center of the ellipsoids by the following construction. Choose x_0 any point. From x_0 proceed in the direction z_1 until the hyperplane P_1 is reached. Designate this point x_1. Now, from x_1 proceed in the direction z_2 until the hyperplane P_2 is reached. Designate this point x_2. Proceed in this manner, getting x_{i+1} from x_i by proceeding in the direction z_{i+1} until the hyperplane P_{i+1} is reached. One easily proves, by induction on i, that x_i is on each of $P_1, \ldots, P_i$. However, the only point common to each of $\mathrm{P}_1, \ldots, P_n$ is the center of the ellipsoids. So x_n is this center.

3. *Algebraic method of finding the center.* Let c be a point and C a real symmetric matrix. Then, for positive real values of ϕ, the surfaces

$$(x-c)^TC(x-c)=\phi \tag{1}$$

form a family of ellipsoids such as considered above, with the common center, c. Indeed, each such family of ellipsoids can be so represented.

If then z_i is a direction, the conjugate hyperplane P_i consists of all points w, such that

$$(w-c)^TCz_i=0. \tag{2}$$

In particular, a direction z_j is parallel to P_i if, and only if,

$$z_j^TCz_i=0 \qquad (i\neq j). \tag{3}$$

We may express (3) in words by saying that z_j and z_i are C-orthogonal. Then if we wish z_i to be parallel to each of $P_1, \ldots, P_{i-1}$, we need only to insure that z_i is C-orthogonal to each of $z_1, \ldots, z_{i-1}$. Since C-orthogonality is a symmetric relation, we see that the condition we wish the $z_1, \ldots, z_n$ to satisfy is that they be C-orthogonal in pairs.

[1] National Bureau of Standards, Los Angeles, California, and Cornell University.

One can find such a set of z's by various means. One scheme is essentially a Gram-Schmidt procedure. Choose $w_1, \ldots, w_n$ a set of independent directions. Now choose $z_1 = w_1$. Then take z_2 to be of the form $w_2 - \alpha z_1$. One readily chooses α so that z_2 is C-orthogonal to z_1. Then take z_3 to be of the form $w_3 - \alpha z_1 - \beta z_2$. One readily chooses α and β so that z_3 is C-orthogonal to each of z_1 and z_2. And so on.

A somewhat simpler scheme is as follows.[2] Take z_1 to be any arbitrary direction. Take z_2 to be of the form $Cz_1 - \alpha z_1$, choosing α so that z_2 is C-orthogonal to z_1. Thereafter, take z_{i+1} to be of the form

$$z_{i+1} = Cz_i - \alpha_i z_i - \beta_i z_{i-1}, \tag{4}$$

choosing α_i and β_i so that z_{i+1} is C-orthogonal to each of z_i and z_{i-1}. Then z_{i+1} is C-orthogonal to each of $z_1, \ldots, z_i$. This is a special case of the following theorem.

Theorem 1. *Let B and C be symmetric matrices which commute. Let $z_0 = 0$, z_1 be arbitrary and*

$$z_{i+1} = Cz_i - \alpha_i z_i - \beta_i z_{i-1}$$

for $i \geq 1$, with α_i and β_i chosen so that z_{i+1} is B-orthogonal to each of z_i and z_{i-1}. Then z_{i+1} is B-orthogonal to each of $z_0, z_1, \ldots, z_i$.

Proof by induction on i. Since $z_0 = 0$, it is B-orthogonal to all z_i. So our conclusion holds for $i = 0$. Let us assume the conclusion for i. By our choice of α_i and β_i, z_{i+2} is B-orthogonal to each of z_{i+1} and z_i. It is also B-orthogonal to z_0. Consider $z_j^T B z_{i+2}$, with $0 < j < i$. By the hypothesis of our induction, z_j is B-orthogonal to each of z_{i+1} and z_i. So by (4) and the assumption that B and C are symmetric matrices which commute,

$$z_j^T B z_{i+2} = z_j^T B C z_{i+1} = z_j^T C B z_{i+1} = z_{i+1}^T B C z_j.$$

By (4) again,

$$z_j^T B z_{i+2} = z_{i+1}^T B z_{j+1} + \alpha_j z_{i+1}^T B z_j + \beta_j z_{i+1}^T B z_{j-1} = 0.$$

Thus z_j is B-orthogonal to z_{i+2}, and our theorem is proved.

In theory, this scheme can terminate by having $z_i = 0$ for $i \leq n$. In a typical computation, there is a zero probability of picking a z_1 that causes the series to terminate too soon, and we may ignore the possibility. In the application to solving simultaneous equations given below, one would have the solution whenever one gets to $z_i = 0$, and so the choice of an initial z_1 is truly immaterial.

Once one has the z's, one easily gets the x's. For, given x_i, one can choose ϕ in (1) so as to get the ellipsoid of our family which passes through x_i. Now x_{i+1} is halfway across this ellipsoid in the direction z_{i+1}.

4. *Solution of simultaneous equations.* Let A be a nonsingular matrix and k a vector, and suppose we wish to solve

$$Ax = k. \tag{5}$$

Take H any real symmetric matrix. Then for real positive ϕ

$$(Ax - k)^T H (Ax - k) = \phi \tag{6}$$

is one of a family of similar, similarly oriented ellipsoids with center $A^{-1}k$. That is, we can write (6) in the form

$$(x - A^{-1}k)^T A^T H A (x - A^{-1}k) = \phi.$$

Then, comparing with (1), we have $c = A^{-1}k$, $\quad C = A^T H A$.

Then the methods outlined in the preceding section suffice to determine the center of (6), which is identical with the solution of (5).

In the present case, we can determine x_n directly by the following device. By our method of successive generation of the x's, we see that x_i has the form $x_0 + a_1 z_1 + \ldots + a_i z_i$. Recalling that x_n is $A^{-1}k$, we see that we wish values of a_i such that

[2] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Research NBS **45**, 255–282 (1950); RP 2133 (see section VII).

$$A^{-1}k=x_0+a_1z_1+. \; . \; .+a_nz_n.$$

Multiplying on the left by $z_j^TA^THA$, and recalling that the z's are C-orthogonal (with $C=A^THA$), e get

$$z_j^TA^THk=z_j^TA^THAx_0+a_jz_j^TA^THAz_j, \tag{7}$$

hich determines a_j.

One should note that for the above procedure, one requires only a knowledge of A^TH, A^THA, k, z_1, id x_0, of which H, z_1, and x_0 are at our disposal. In case A is symmetric and positive definite, one can ke $H=A^{-1}$, with considerable simplification of the computations. Otherwise, $H=I$ is a convenient ıoice.

5. *First iterative scheme.* If one has any set of z's that are C-orthogonal, it clearly suffices to replace ıem by any multiples of them. Thus, we can generalize (4) to

$$z_{i+1}=\gamma_i(Cz_i-\alpha_iz_i-\beta_iz_{i-1}). \tag{8}$$

ı case A is symmetric and positive definite, we take $H=A^{-1}$, so that $C=A$. Then the following scheme oceeds according to (8).

We define the residuals

$$r_i=k-Ax_i, \qquad (0\leq i). \tag{9}$$

We note that our scheme for generating the x's is expressed by

$$x_i=x_{i-1}+a_iz_i, \qquad (1\leq i), \tag{10}$$

here

$$a_i=\frac{z_i^Tr_{i-1}}{z_i^TAz_i}, \qquad (1\leq i). \tag{11}$$

y (9) and (10),

$$r_i=r_{i-1}-a_iAz_i, \qquad (1\leq i). \tag{12}$$

By use of this, one can simplify (8) to

$$z_{i+1}=r_i+b_iz_i, \qquad (0\leq i), \tag{13}$$

here

$$b_0=0, \tag{14}$$

$$b_i=-\frac{z_i^TAr_i}{z_i^TAz_i}, \qquad (1\leq i). \tag{15}$$

Clearly, by (11) and (12)

$$z_i^Tr_i=0, \qquad (1\leq i). \tag{16}$$

lso by (13) and (15),

$$z_i^TAz_{i+1}=0, \qquad (1\leq i). \tag{17}$$

so by (13) and (14),

$$z_1=r_0. \tag{18}$$

By (13) and (16)

$$r_i^Tz_{i+1}=|r_i|^2. \tag{19}$$

Theorem 2. If $r_i\neq0$ for $0\leq i<m$, then for $1\leq j\leq m$, $z_j\neq0$, so that a_j and b_j can be defined by (11) *d* (15).

Proof. By (13) and (16)

$$|z_{i+1}|^2=|r_i|^2+b_i^2|z_i|^2. \tag{20}$$

From this, our theorem follows by induction on m.

Clearly, if $r_i \neq 0$ for $0 \leq i < m$ and $r_m = 0$, then by (9), $x_m = A^{-1}k$ and we have solved (5). Hence one would never wish to carry this procedure past the least m for which $r_m = 0$.

Theorem 3. *If* $r_i \neq 0$ *for* $0 \leq i < m$, *then for* $1 \leq j \leq m$, $a_j > 0$.

For by (19) and (11).

$$a_j = \frac{|r_{j-1}|^2}{z_j^T A z_j}. \tag{21}$$

But A is positive definite, and $z_j \neq 0$ by theorem 2, so that $z_j^T A z_j > 0$.

Theorem 4. *If* $r_i \neq 0$ *for* $0 \leq i < m$, *then*

$$z_i^T r_j = 0 \qquad (1 \leq i \leq j \leq m) \tag{22}$$

$$z_j^T r_i = |r_{j-1}|^2 \qquad (0 \leq i < j \leq m) \tag{23}$$

$$z_i^T A r_j = z_i^T A z_{j+1} \quad (1 \leq i \leq m, \quad 0 \leq j \leq m, \quad i \neq j) \tag{24}$$

$$r_i^T r_j = 0 \qquad (0 \leq i < j \leq m) \tag{25}$$

$$z_i^T A z_j = 0 \qquad (1 \leq i < j \leq m+1). \tag{26}$$

Proof by induction on m. First let $m=1$. Then we get (22) from (16), (23) from (19), (24) from (18), (25) from (16) and (18), and (26) from (17).

Assume the theorem for m, with $m \geq 1$. Further assume that $r_i \neq 0$ for $0 \leq i < m+1$. By (12)

$$z_h^T r_i = z_h^T r_{i-1} - a_i z_h^T A z_i.$$

So by (26), for $1 \leq i \leq m+1$, $1 \leq h \leq m+1$, $i \neq h$, we get $z_h^T r_i = z_h^T r_{i-1}$. If we combine this with (16), we infer (22) for $m+1$. If we combine this with (19), we infer (23) for $m+1$.

By (13), we get

$$z_i^T A z_{j+1} = z_i^T A r_j + b_j z_i^T A z_j.$$

From this, we get (24) for $m+1$ by using (14) when $j=0$ and (26) when $1 \leq j$.

By (12)

$$r_i^T r_{m+1} = r_i^T r_m - a_{m+1} r_i^T A z_{m+1}.$$

If $i < m$, then $r_i^T r_m = 0$ by (25), $r_i^T A z_{m+1} = z_{m+1}^T A z_{i+1}$ by (24), and $z_{m+1}^T A z_{i+1} = 0$ by (26). Also, i $i=m$, then we get

$$r_m^T r_{m+1} = |r_m|^2 - a_{m+1} r_m^T A z_{m+1}.$$

By (19) and (24)

$$r_m^T r_{m+1} = z_{m+1}^T r_m - a_{m+1} z_{m+1}^T A z_{m+1}.$$

Then by (11), $r_m^T r_{m+1} = 0$. Thus we infer (25) for $m+1$.

By (17) $z_{m+1}^T A z_{m+2} = 0$. By (13) and (26), we have for $1 \leq i \leq m$, $z_i^T A z_{m+2} = z_i^T A r_{m+1} = r_{m+1}^T A z_i$ But by (12) and theorem 3

$$A z_i = -\frac{1}{a_i} r_i + \frac{1}{a_i} r_{i-1}$$

So

$$z_i^T A z_{m+2} = -\frac{1}{a_i} r_{m+1}^T r_i + \frac{1}{a_1} r_{m+1}^T r_{i-1}.$$

Then by (25), $z_i^T A z_{m+2} = 0$. Thus we infer (26) for $m+1$.

We note that (26) tells us that the z's are A-orthogonal, so that we have indeed a special case o the procedures outlined above.

Theorem 5. *If* $r_i \neq 0$ *for* $0 \leq i < m$, *then for* $1 \leq j \leq m$, $b_j \geq 0$.

Proof. By (12) and (25),

$$|r_j|^2 = -a_j r_j^T A z_j. \tag{27}$$

en by (21) and (15),

$$b_j=\frac{|r_j|^2}{|r_{j-1}|^2}. \tag{28}$$

From (28), we see that if $r_i\neq 0$ for $0\leq i<m$ and $r_m=0$, then $b_m=0$. Then by (13), $z_{m+1}=0$.

It is interesting that the iteration outlined above is identical with that derived by Stiefel from an irely different approach.[3]

By (12) and (13),

$$z_{i+1}=r_{i-1}+b_iz_i-a_iAz_i$$

$$r_{i-1}=z_i-b_{i-1}z_{i-1}.$$

$$z_{i+1}=-a_iAz_i+(1+b_i)z_i-b_{i-1}z_{i-1}. \tag{29}$$

ence the z's are indeed constructed according to (8).

Not only is x_n the solution of (5), but one can now readily solve $Ay=h$. To do so, let us seek c's hat $A^{-1}h=c_1z_1+\ \ldots\ +c_nz_n$. That is, $h=c_1Az_1+\ \ldots\ +c_nAz_n$. So, since the z's are A-orthogonal, $=c_jz_j^TAz_j$. Thus

$$A^{-1}h=\sum_{j=1}^{n}\frac{z_j^Th}{z_j^TAz_j}z_j$$

$$=\left(\sum_{j=1}^{n}\frac{z_jz_j^T}{z_j^TAz_j}\right)h.$$

ting h be an arbitrary vector, we infer

$$A^{-1}=\sum_{j=1}^{n}\frac{z_jz_j^T}{z_j^TAz_j}.$$

One can proceed entirely by means of the r's and x's without ever bringing in the z's at all. For by 1) and (13),

$$r_{i+1}=r_i-a_{i+1}Az_{i+1}$$

$$r_i=r_{i-1}-a_iAz_i$$

$$Az_{i+1}=Ar_i+b_iAz_i.$$

$$r_{i+1}=-a_{i+1}Ar_i+\frac{a_i+a_{i+1}b_i}{a_i}r_i-\frac{a_{i+1}b_i}{a_i}r_{i-1}.$$

Similarly, from (10) and (13), we get

$$x_{i+1}=a_{i+1}r_i+\frac{a_i+a_{i+1}b_i}{a_i}x_i-\frac{a_{i+1}b_i}{a_i}x_{i-1}.$$

Thus we can generate the x's and r's by the following recursion.

Choose x_0 at random. Define $r_0=k-Ax_0$, see (9). Define $x_1=x_0+\alpha_0r_0$ (see (10) and (18)). Then (9) $r_1=r_0-\alpha_0Ar_0$. Hence we are able to determine α_0 from the condition $r_1^Tr_0=0$ (see (25)).

One can now proceed by setting

$$x_{i+1}=\alpha_ir_i+\beta_ix_i+(1-\beta_i)x_{i-1}.$$

en by (9)

$$r_{i+1}=-\alpha_iAr_i+\beta_ir_i+(1-\beta_i)r_{i-1}.$$

nce one can determine α_i and β_i from the conditions $r_{i+1}^Tr_i=0$ and $r_{i+1}^Tr_{i-1}=0$.

6. *Other iterative schemes.* Since theorem 1 holds for any B which commutes with C, we can generize the scheme presented at the end of the last section by using an arbitrary B. Then the scheme ears as follows.

E. Stiefel, see paper 5, page 43.

We have the symmetric matrix A (the assumption of definiteness can now be relaxed) and the vecto k. Choose B any symmetric positive definite matrix which commutes with A. For example, B coul be I or $\alpha I+A$, where α is chosen large enough to insure definiteness. We now undertake to generate series of approximations $x_0, x_1, \ldots$ to $A^{-1}k$, with the property that the residuals $r_i=k-Ax_i$ shall b B-orthogonal. We merely follow the scheme outlined above. We choose x_0 at random. Then w choose α_0, so that r_0 is B-orthogonal to $r_0-\alpha_0Ar_0$. Then we define $x_1=x_0+\alpha_0r_0$, whence by (9 $r_1=r_0-\alpha_0Ar_0$. In general, we choose α_i and β_i, so that

$$-\alpha_iAr_i+\beta_ir_i+(1-\beta_i)r_{i-1}$$

shall be B-orthogonal to r_i and r_{i-1}. Then we define

$$x_{i+1}=\alpha_ir_i+\beta_ix_i+(1-\beta_i)x_{i-1},$$

whence by (9)

$$r_{i+1}=-\alpha_iAr_i+\beta_ir_i+(1-\beta_i)r_{i-1}.$$

It may be more efficient to compute the increments of the x's by the formula

$$x_{i+1}-x_i=\alpha_ir_i-(1-\beta_i)(x_i-x_{i-1}).$$

By theorem 1, the r's will be B-orthogonal in pairs. Consequently, there must be an m, $0\leq m\leq$ such that $r_m=0$ (and so $x_m=A^{-1}k$) unless the procedure terminates sooner because our attempt to dete mine α_i or β_i requires a division by zero. In the preceding section, we showed that if A is positive definit and B is the identity matrix, a division by zero could not be called for before one found an $r_m=0$. I the general case, it is not known what can happen, but it seems plausible that one will have little dange in general of encountering a division by zero.

7. *Rounding off errors.* In the scheme of section 6, one always computes the residuals directl from A, k, and x_i. Hence there is no danger of accumulating a round-off error in the sense that on thinks one has a good x_i, when, in fact, one has a poor x_i because of round-off errors. One may indee have a poor x_i because of round-off errors (or even because of actual mistakes) when one should have good x_i, but the poorness of x_i will be fully apparent.

If there were no round-off (or other) errors, the procedure would surely give $A^{-1}k$ in n or fewe steps. However, because of round off, the r's will not be truly B-orthogonal, and so the process will no terminate in n steps. However, since it should terminate in n steps, it can be expected to come ve close in n steps.

Note that after the first step, there is nothing to distinguish one step from the next except the siz of the r's. If at the end of n steps, the r_n is not sufficiently small (either from round off or from mi takes), there is nothing whatever to bring the process to a termination. One can simply let the pr cedure continue until r_m is sufficiently small, even if this may in some cases require a value of m great than n.

For the procedure of section 5, similar remarks hold. If x_i is computed from (10) and then is computed from (9), leaving (12) as a check, one is dealing with actual residuals, and is protecte from the possibility of an error slipping in undetected. Also, there is absolutely nothing to terminat the procedure at the nth step if by some mischance r_n is not sufficiently small.

It thus appears that in actual computation, our schemes are iterative schemes which will no ordinarily terminate in any finite number of steps, but which may be expected to give a very clos approximation in n steps.

Nothing is yet known about the rapidity of convergence after n steps, but it seems plausible t assume that it is comparable to the rapidity of convergence during the first n steps. Investigation this question is called for, and a numerical testing of the procedures is now in progress. Another que tion of a related sort is that of what conditions on the matrix A and the vector k make it likely th round-off errors could force r_n to be larger than allowable, and whether, in such case, it is more profitab to let the iteration continue or to make a new start with the x_n which has been obtained.

8. Some Problems in Aerodynamics and Structural Engineering Related to Eigenvalues

R. A. Frazer [1]

Scope of the Paper

The problems of aerodynamics and structural engineering with which this paper is concerned relate moving fluids and elastic bodies and therefore to continuous systems. The analysis of the effects small disturbances of such systems leads to sets of simultaneous linear and homogeneous equations ich contain one or more parameters. Without attempting a rigorous or general definition, I shall erpret the term "eigenvalue" to mean any value of a typical parameter for which such sets of equations ve a definite nontrivial solution. Other terms in current usage are "proper value" and "characteristic mber".

It will be obvious that, in general, the only practicable method of solution of these problems y means of approximations. Even so, an approximate treatment of the continuous systems concerned uld present prohibitive difficulties were it not for one circumstance. This is bound up with the ctical uses to which the solutions are finally to be put. An aircraft designer concerned with these blems is not interested to know how his structure would behave in all conceivable circumstances; only wants to be assured that it will be safe in actual use. To the mathematician, this limitation ans that only a few of the infinite number of possible sets of eigenvalues need be determined, and takes advantage of this fact to simplify his analysis. His procedure, in brief, is to replace the conuous system by a finite one chosen in such a way that the two systems have (approximately at least) appropriate band of eigenvalues in common. To obtain successful approximating systems, he has, present, to rely heavily on physical intuition. A more abstract treatment may be found possible en the general principles underlying the approximations are better understood. But, for the purposes this paper, many references to the physical background will be necessary in order to provide a clear ount.

Approximating Systems in Flutter Analysis

One problem of great practical importance in relation to airplane design is the prevention of unble structural oscillations of the type known as flutter. With a defective design, it is possible for airplane in steady flight at any given altitude to develop flutter for certain ranges of the flight speed. e analysis is generally restricted to a calculation of the "critical speeds", which define the end points the speed ranges for flutter and which therefore lead to simple harmonic oscillations of the structure. knowledge of the lowest critical speed is usually sufficient for practical purposes, as this decides the ghest safe speed for the airplane at the given height.

The essential task confronting the flutter analyst, then, is not to describe fully the flutter properties a given airplane, but merely to locate its lowest critical speed. Moreover, great precision in the termination of this speed will be unnecessary because a factor of safety will be allowed. He is theree entitled to substitute any mechanically different and simpler system which possesses approximately e same lowest critical speed. A replacement of this nature, to give the same accuracy, could obviously done in many different ways; and the optimum way in any given case would depend partly on the ount of calculation involved and partly on the ease of formulation. It is probable that there is no e procedure that is optimum for all classes of flutter problems. At present there is no certainty even to the best methods to be used for particular classes.

[1] National Physical Laboratory at Teddington, Middlesex, England. The paper is published by permission of the Director of the Laboratory.

The first step in a flutter analysis is to decide on the type of the approximating system and on the number of the freedoms. A description of the most generally favoured system will serve as an illustration; some others will be mentioned later. The orthodox system is a semirigid copy of the airplane, or of the aircraft component concerned. This system has the same external appearance and the same mass distribution as the original structure, but is a mechanism possessing chosen deformation modes in its several freedoms. It is usual to base these modes and freedoms on the lower order natural oscillations of the structure, and these are sometimes calculated and sometimes determined experimentally by resonance tests. This procedure is probably the safest, but satisfactory results have also been obtained from modes appropriate to static loading, or from modes represented by polynomials or other simple functions of position. The number of modes and freedoms to be included in a flutter calculation is usually a matter for judgment. With complicated problems, such as those involving possible interactions between the wings, tail organs and fuselage, a decision may be very difficult. In doubtful cases, it may well be necessary to introduce more and more freedoms until results have settled down. The greatest number of freedoms used for a single systematic flutter analysis at Teddington has been twelve. The investigation related to the natural oscillations and critical flutter speeds of an exceptionally large airplane [1], and the numerical work, which was mainly carried out by two computers using electrical machines, occupied about two years.

Once the approximating system has been chosen, the next step is to obtain the expressions for the aerodynamic loads on the oscillating system. A knowledge of these aerodynamic coefficients is vital to the whole analysis, and much effort is being devoted to their determination both theoretically and experimentally. However, this highly important and highly specialised branch of aerodynamics can only be mentioned here in passing. To complete the analytic formulation, the appropriate nondimensional aerodynamic, inertial, and elastic coefficients are introduced into the dynamical equations. If the assumed system has n degrees of freedom, two conjugate complex sets of n simultaneous algebraic equations are finally obtained, each set being linear and homogeneous in n complex unknowns. These unknowns determine the amplitude and phase relations between the oscillatory movements in the different freedoms.

The parameters dependent on the flight conditions which can be present in the equations are the speed parameter (which is inversely proportional to the square of the flight speed), the frequency parameter (which is proportional to the ratio of frequency to flight speed), the Mach number (ratio of flight speed to speed of sound in the free stream), and the air density ratio (flying height to ground level). All these must take real positive eigenvalues, and some are interrelated. Other parameters will also be present if any of the data connected with the actual design of the structure are variable.

The critical equations can be expressed by matrices as

$$[u\pm iv]\{x\pm iy\}=0, \tag{1}$$

where u and v denote real square matrices of order $n,x\pm iy$ denote the column vectors of the unknowns, and i denotes the imaginary unit. Then the compatibility conditions in determinantal form are

$$\Delta_1\equiv|u+iv|=0 \quad \text{with} \quad \Delta_2\equiv|u-iv|=0. \tag{2}$$

The matrices u and v will in general be asymmetric owing to the asymmetry of the aerodynamic coupling coefficients. Their elements will be linear in the speed parameter but transcendental in the frequency parameter and Mach number. Thus, in general, the only practicable basis of calculation is to introduce trial values for the frequency parameter and the Mach number until both conditions (2) are satisfied by a corresponding real positive value for the speed parameter. This process can sometimes be assisted by a judicious use of any structural design parameters present. It is also greatly simplified when the effects of compressibility can be ignored and the Mach number is thus absent.

In British practice, the transcendental functions of the frequency parameter in the aerodynamic coefficients relevant to normal flight speeds are sometimes replaced by quadratic approximations. The expanded forms of the determinantal conditions (2) can then be interpreted graphically as two polynomial curves of degrees n and $n-1$ in the plane of the speed and frequency parameters. Corresponding pairs of practical eigenvalues are given by those intersections of the curves that lie in the first udrant. Several methods of calculation based on interpolation formulas for the polynomials have en suggested [2, 3]. In one of these, an interpolation framework of straight lines $L_1=0$, $L_2=0$, $L_3=0$, n so on, is adopted, no two lines being parallel and no three concurrent. The polynomials are then anded in bivariate partial fractions in which the denominators are the products of the linear expres-cs L_1, L_2, L_3, etc., taken in pairs. The numerators of the fractions are the values of the polynomials the line intersections, and these are calculated directly as numerical determinants Δ_1 and Δ_2. With i use of a standard line framework and prepared double-entry tables for the fractions, the search for intersections of the curves becomes a routine procedure that might perhaps be suitable for Hollerith ther automatic computing equipment. A disadvantage of these methods is that with many degrees reedom they demand very high accuracy.

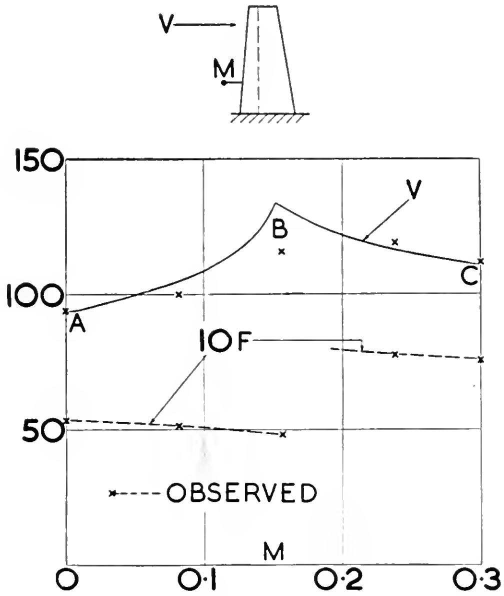


FIGURE 8.1. *Critical speeds V and frequencies F for model wing.*

An example of a simple flutter diagram is given by figure 8.1. The diagram compares the lowest ical speeds observed for a model cantilever wing in a wind tunnel [4] with results calculated by the ariate partial fractions method. The model wing carried a variable concentrated mass M representing ing engine, and the approximating system adopted had four freedoms and simple polynomial distortion des. Standard air density was assumed, and the three parameters present in the analysis were the ical speed, the critical frequency, and the engine mass. The modes of oscillation and the frequencies e markedly different for the two branches AB, BC of the critical-speed curve. For AB the wing tion containing the engine mass showed little movement, but with BC it pitched vigorously and at a iceably higher frequency. When the engine mass was adjusted to accord with the nodal point B, ier type of oscillation could be obtained according to the way in which the wing was disturbed.

An example of an unorthodox approximating system in flutter analysis is provided by a segmented iirigid wing. Systems of this type have been used [5] in connection with the flutter of solid blades l airscrews. The elastic wing is regarded as segmented by crosscuts into either rigid slices, or into es that are simply semirigid so that continuity of the flexural and torsional displacements is preserved r the span. The slices are interconnected by springs that are so adjusted that the fundamental illations of the actual and approximate systems agree. Critical speeds predicted by this method l by the standard semirigid method have been found to agree well.

Approximations by Lumped Coefficients

Another possible means of representing a continuous system approximately is by a system of crete massive particles with elastic connections. The theory of the relationships between continuous and discrete systems is, of course, a very wide subject, which has been treated on a general basis b matrices in recent papers by several writers [6, 7], both in relation to mechanics and electrodynamics

The form of approximation relevant to the present context is the simple process that is sometime used to obtain estimates of the natural frequencies of elastic beams and shafts. It is often referred t by engineers as the method of "lumped masses". The application of this principle to flutter analysi has hitherto been very limited. Partial use of it is made when special parts of an aircraft structure such as engine nacelles, wing fuel tanks or fuselage components are replaced by one or more massiv particles. But fully discrete systems, in which not only the masses or inertial coefficients, but also th aerodynamic coefficients are "lumped", have not so far been systematically studied. A simple exampl of such a system will be given later.

Some instructive conclusions on the effects of mass lumping can be drawn from oscillating tensione strings. The natural frequencies of transverse oscillation of stretched continuous strings and approximating discrete strings have been compared by Den Hartog [8]. He assumes that all the strings hav the same length and same total mass μ, and that each discrete string is massless but carries equal massiv particles at equal distances apart across the span. As the total string mass is kept constant, the mas values for the individual particles will depend on whether or not one of the particles is also allowed t be present at each fixed end. Results are given by Den Hartog for both cases. When the chain o particles stops short of the ends, the approximate fundamental frequencies are too low, but when i includes the ends the frequencies are too high. In either case, the true frequency value is approache only slowly as the number of the particles is increased. It can however, be objected that neither o these arrangements provides a really fair representation of a continuous string. If such a string is t be segmented into links, the simplest procedure evidently is to assign one-half of the mass of each linl to the end points of that link. So, if in particular the string is uniform and the links are all equal, th discrete system carries equal massive particles within its span, but only *half* masses at its two ends On this supposition the method is, in fact, more promising, as will be seen from figure 8.2.

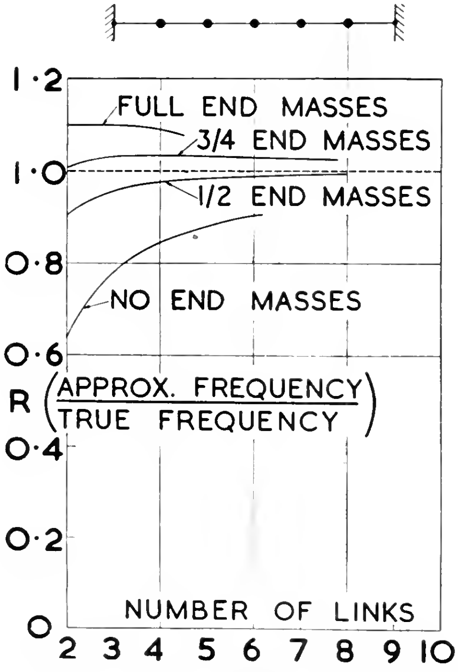


FIGURE 8.2. *Lowest frequencies of discrete strings.*

When half masses are present at the ends, the natural frequencies of any given order corresponding to the discrete and the continuous strings are in the ratio

$$R=\frac{\sin\,(j\pi/2n}{(j\pi/2n)}, \qquad (3)$$

here n denotes the number of equal links in the approximate system and j (which ranges from 1 to -1) denotes the number of antinodes in the oscillation wave form. With the fundamental mode $j=1$) the error in frequency is 10 percent when only two links are used, and is less than 5 percent with aree links. In the case of the first overtone, four links and six links are needed, respectively, for the ıme accuracies.

In connection with this method of approximation, it is to be noted that the differences between ıe displacement modes can be pronounced, even when the frequencies agree well. With the continuous ring, the wave forms of displacement are sine curves. With the discrete strings, the modes are defined ı the first place by the displacements of the lumped particles and in the second place by the displacements f the links. The displacements of the particles accord with sine curves, but the links are massless and emain straight. So the complete modes are composed of blunted sine curves, the blunting increasing s the number of links is reduced. If modes free from these imperfections are required, they must e calculated for the continuous string with the use of the approximate frequencies.

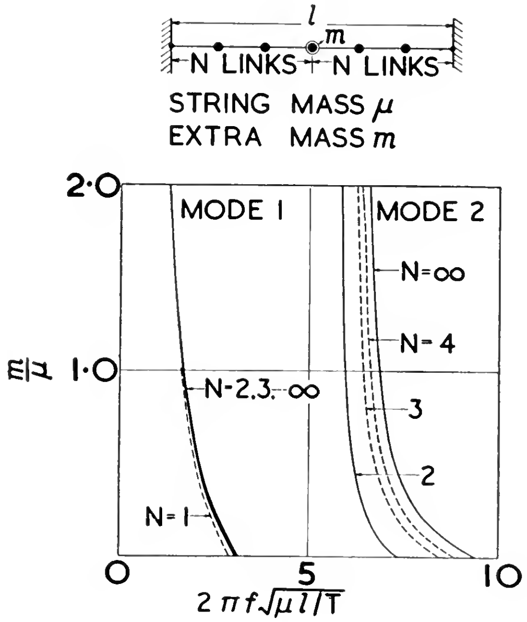


FIGURE 8.3. *Symmetric oscillations of loaded discrete strings.*

Figure 8.3 gives some comparisons of natural frequencies for continuous and discrete strings of qual length, l, and equal mass, μ, when an extra concentrated mass, m, is carried at the centre. In ach case the string mass is made up of equal equidistant particles except that only half-particles are laced at the ends. The ordinate in the diagram is the ratio m/μ (extra mass to string mass), and the bscissa is proportional to the natural frequency. For simplicity, results are only shown for the fundaıental and first overtone symmetric oscillations. It is seen that the discrete strings all provide good rst approximations for the fundamental frequency, and that fair approximations are obtained in the ase of the first overtone even when only few particles are used.

A similar method of approximation is possible for more general systems containing two or more ensioned strings cross connected by sets of rigid rods or ribs. A simple analogue to a uniform rectangular irplane wing with its two spars held at both ends can be devised by substituting two tensioned massless trings for the wing spars and cross connecting them by a close pack of separate thin ribs. If the number f rib strips is very great, this articulated imitation wing is effectively continuous. Any natural transerse oscillation will, in general, consist of a torsional movement about some axis parallel to the span, nd the twist will be distributed sinusoidally across the span.

To obtain an approximating system, it is only necessary to segment the strings into equal finite nks and to treat the rib masses as though they were concentrated at the ends of the links. The wing cover can be pictured as being carried by the ends of the links in solidified pieces, the divisions between the pieces being central in the links. Thus, when the system oscillates, each discrete strip of wing moves with its bridging rib two-dimensionally, pitching angularly and translating vertically. The two half strips attached to the fixed ends will, of course, remain idle. As might be expected, corresponding natural frequencies of the approximating and continuous wing systems are here in the same ratio as that found for a single string.

If the continuous wing had unfavorable inertias and elastic stiffnesses and were exposed to steady winds, it would be liable to flutter in modes involving coupled bending and torsion. This simple flutter problem can be solved formally and in finite terms, on the assumption that aerodynamic coefficients appropriate to two-dimensional air flow are applicable over the wing surface. The problem is also soluble for the approximating systems, provided the coefficients used to evaluate the aerodynamic forces over the discrete strips of wing are again applicable to two-dimensional air flow. The procedure in the approximate calculation, therefore, is, effectively, to lump not only the masses but also the aerodynamic coefficients at the ends of the links. It is found that for incompressible flow corresponding critical speeds for the approximate and continuous wings are again in the same ratio as that applicable for natural frequencies. In particular, the lowest critical speed (namely, that appropriate to flutter in the first mode) will be 10 percent in error with only two links or one active strip, and will be under 5 percent in error with three links or two active strips. These approximations would be too low and would thus err on the safe side if practical applications were in view. All the wings will show flutter in any given mode for the same value of the frequency parameter. These conclusions can be drawn without any specific knowledge of the way in which the two-dimensional aerodynamic coefficients depend functionally on the frequency parameter. Moreover, they would not be influenced very materially by the effects of compressibility.

The example just given is crude, and it may well be misleading. However, there is some other evidence that approximations of this type deserve more attention. The well-known method due to Theodorsen and Garrick [9] for the prediction of the flutter of cantilever wings is effectively dependent on lumped coefficients, and it gives results that compare well with those obtained by more elaborate methods [10]. In general, a lumped coefficient method of flutter prediction would have several advantages and one main disadvantage. On the one hand, it would require no previous knowledge of the distortion modes and would be relatively easy to formulate. On the other hand, if the convergence were slow, the method would prove expensive in regard to the number of freedoms, because each additional link introduced would require a corresponding increased number of freedoms. More light on this question and on the general flutter properties of actual wings may be obtained from the study of further simple articulated wing systems, such as those carrying extra concentrations of mass or having more general plan forms. Another obscure and vital question is the possible application of lumped coefficients to control surfaces.

Lumped Coefficients in Structural Engineering

Before some of the other problems of fluid dynamics are mentioned, it may be of interest to exemplify the use of lumped coefficients in structural engineering. The natural oscillations of framed structures were first treated in detail by Reissner [11]. He regards the structure as a dynamical system composed of interconnected massless elastic bars with compensating lumped masses placed at all the joints. However, the complications connected with the many degrees of freedom and with the solution of the frequency equation precludes the general use of his method for extensive structures such as bridges.

A theoretical study of the stiffness and frequency properties of bridge box structures [12] has recently been carried out at Teddington as part of an investigation on the aerodynamic oscillations of suspension bridges, and some of the results obtained throw further light on the influence of lumped masses on natural frequencies. Before the frequency diagrams are shown, some explanations on the basis of calculation will be necessary.

A suspension-bridge box structure consists of a number of identical framed box-form bays connected in series by cross girders. These cross girders play, effectively, the part of transverse diaphragms separating the bays. In practice the weight of the structure is taken by the cables through vertical ropes, or suspenders, attached to the cross girders. However, the present method of analysis is concerned olely with the frequency properties of the suspended structure itself in the absence of suspenders and a ravity field. As an approximation, the masses of the structure are assumed to be lumped, not at the ndividual joints as in Reissner's theory, but at cross-girder positions; and the inertial condition is decribed as "close" or "open" according as all the cross girders or merely a selected equidistant set are ass-loaded. The treatment thus resembles that already explained for strings, except that with the ridge the links consist of lengths of the elastic structure which are necessarily multiples of the span of a ingle bay.

When the structure bends vertically any cross girder is assumed to displace as a rigid body, and it is llowed two degrees of freedom—vertical translation and angular flexural displacement. With *open* ading the displacements of the mass-loaded cross girders are regarded as primary. The other girders e within the lengths of structure that form the links, and their displacements are therefore determined y the primary displacements and the elastic stiffness constants of the system. When the structure scillates, the links do not remain straight as for strings, but curve.

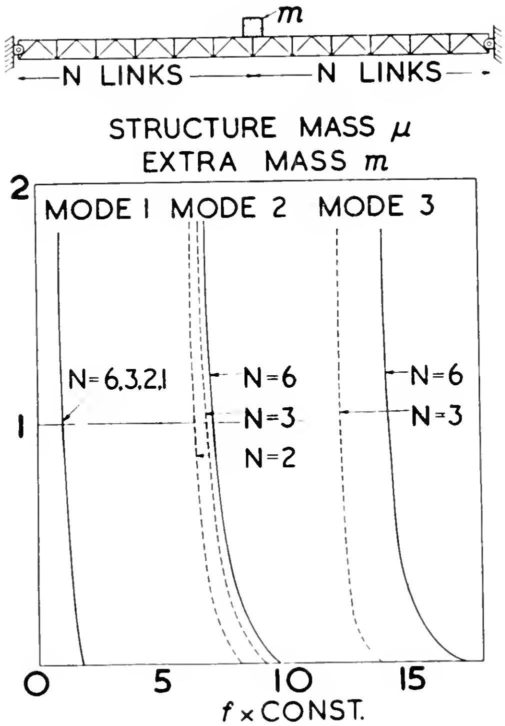


FIGURE 8.4. *Bending oscillations of a bridge box structure.*

Figure 8.4, which is plotted similarly to figure 8.3, illustrates the influence of open loading on the ymmetric flexural frequencies of a box structure of total mass μ, with an extra concentrated mass m t midspan. The datum system is the close-loaded structure containing six simple bays in each half-span, nd the most open one contains merely one link in each half-span. Only the three lowest order oscillaons are covered by the diagram. The system with only one link in each half-span yields a single scillation, but it nevertheless provides a good approximation for the fundamental frequency.

The corresponding results for symmetric torsional oscillations are shown in figure 8.5. Here the ssumed freedoms are the twist of the structure at the cross-girder positions, and the twisting or warping isplacements of the cross-girder planes.

Both of the two problems that have been exemplified are formally soluble, on the assumptions made, n finite terms. The dynamical equations determining the movements of the mass-loaded cross girders nder the elastic forces due to the structural links yield a set of simultaneous recurrence relations connectg the displacement amplitudes of any three such consecutive girders and containing the squared freuency as a parameter. A treatment of these relations by the methods available for simultaneous difference equations leads to a formal solution. An alternative method is to express the equations in bulk by matrices and so to relate the problem to the determination of latent roots. Problems of this type yield partitioned matrices composed of alternant submatrices and irregular rows dependent on the extra mass-concentrations present. The examples given are, of course, only special cases of more general repetitive systems or dynamic chains, composed of any number of identical massive semirigid bodies or carriers joined together in series by identical massless elastic connecting sytems.

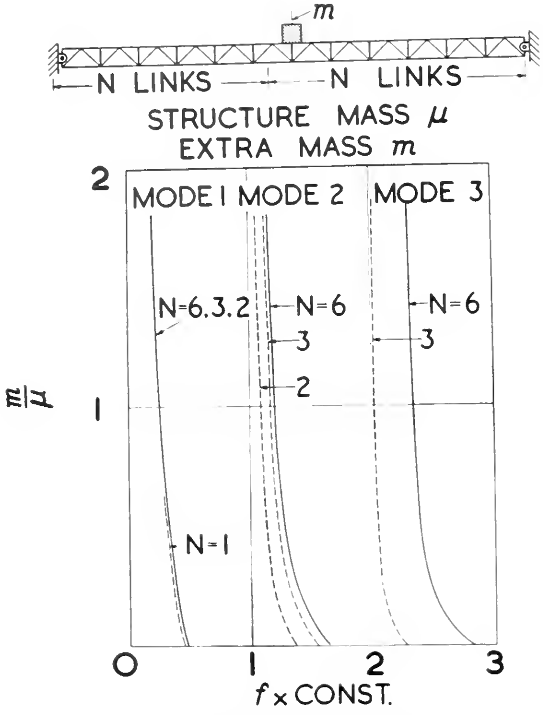


FIGURE 8.5. *Torsional oscillations of a bridge box structure.*

The results for the box structure suggest that it may also be possible to obtain satisfactory estimate of the lower natural frequencies of complete suspension bridges by means of lumped coefficients. However, no definite conclusions on this important question can be drawn at present. A knowledge of thes frequencies plays a vital part in the prediction of the critical speeds at which oscillations of the bridg are set up due to wind. With a long bridge the normal procedure in frequency calculations is to regar the framed suspended structure in each span of the bridge as effectively a continuous elastic beam and t replace the discrete sets of vertical suspenders by continuous sheets of suspenders. The displacement modes of the structure and cables are then given by simultaneous linear differential equations containing the squared frequency as a parameter and coefficients which are variable with distance along the span The analytical difficulties presented by these equations are usually countered by further simplifying assumptions. It is seen, then, that the conventional calculation smoothes out the panneling of the actua bridge design and regards the distribution of the loading to be continuous. The proposed alternative treatment is, in effect, to reduce the number of effective load-bearing cross girders and suspenders to a few only. However, this method requires careful investigation because with systems such as suspension bridges, which are controlled partly by gravity, mass lumping is accompanied by a corresponding lumping of the gravitational stiffnesses. The variation of these stiffnesses with the distance along the span is another complication. It has been suggested by Pugsley [13] that the use of only a few flexibility coefficients, determined for the cables and suspended structure separately, might prove adequate for a suspension-bridge analysis.

Stability of Fluid Flow

To conclude this short paper, a few remarks may be added on one or two of the other aerodynamic roblems that are related to eigenvalues or simultaneous equations.

One subject that has tried the ingenuity of mathematicians for many years is the stability of the ow of an incompressible viscous fluid. The first problem of this type to be attacked successfully conerned the flow between two coaxial circular cylinders in steady rotation. It has been shown by Taylor 4], both experimentally and theoretically, that when the rates of rotation and the radii are appropriately elated the simple two-dimensional rotary flow between the cylinders becomes divergently unstable. n the critical, or neutral, state the flow pattern assumes a banded appearance due to alternations of ositive and negative vortices spaced regularly along the axis. Taylor bases his analysis on Fourier-essel expansions, and the eigenvalues he determines are the widths of the vortex bands. He calculates hem by the approximate solution of a determinantal equation of infinite order.

An alternative treatment of the same problem proposed by Jones [15] shows the possibilities of mple methods of approximation. He represents the velocity components of the flow by polynomials hat satisfy all the boundary conditions and contain sets of free constants. The errors, when these xpressions are substituted in the flow equations, are made zero for a certain number of equidistant adii and the disposable constants are eliminated. This simple collocation process leads to a determinantal equation of finite order for the critical wave length, the order depending on the number of constants sed. In the cases tried out numerically, good agreement with Taylor's results was obtained.

With more general boundary shapes, approximations must also be used to determine the steady ndisturbed flow, and this in itself presents a severe problem. Recent papers by Chi-Teh Wang and rodsky [16, 17] have shown that solutions of fluid-flow problems can be obtained successfully by the ayleigh-Ritz procedure or by Galerkin's method. However, a direct application of such methods to he complete hydrodynamical equations leads to systems of nonlinear simultaneous equations for the ree constants. To overcome this difficulty it is usual to approach the solution by successive linear pproximations.

One such method that has been suggested recently [18] for incompressible flow assumes that the roblem has been solved for a perfect fluid, so that the velocity, q, with which the fluid slips tangentially t each point of the solid boundary is known. The body is then imagined to be covered with a thin xtensible skin which is moved continuously along the surface with speed $(1-\lambda)q$ at each point, where λ s a variable parameter. If viscosity is now introduced, a whole class of problems is obtained, all relating to the same boundary shape and the same viscous stream and differing only as regards the condition f slip at the boundary. The problems range from $\lambda=0$ (when full slip is allowed) to $\lambda=1$ when the slip s completely annihilated. If the solution for the class is assumed to be expansible in ascending powers f λ, the terms are given in succession by a sequence of linear differential equations. The first term of he expansion is the known solution for perfect flow, and the second term is given by the set of differential equations treated by numerical methods by Southwell and Squire [19]. A build-up of an approximate solution by the methods of Galerkin or collocation would be attractive from the standpoint of enerality, but would require equipment for the solution of linear numerical equations of very large rder. This method has not yet been put to a test, and its effectiveness can at present only be conjectured.

The study of disturbed unsteady flow presents unusual difficulties and is an almost unexplored ield. Perhaps the simplest illustration of a problem of this type is shown by the behavior of fluid ontained within a long rotating glass tube. If, when both the tube and the fluid are in steady rotation ogether, the tube is abruptly arrested, it will be found that the fluid motion does not decay two-dimensionally but that it develops vortex bands similar to those observed by Taylor with coaxial cylinders. The peculiarity of this problem is that the deviant flow due to small disturbances becomes significant n relation to the undisturbed flow, although both motions eventually die away. No definition of nstability appears to exist which covers phenomenon of this kind.

The subject of disturbed unsteady motion has wide applications and includes for instance the tability analysis of steady aerodynamic oscillations. Such problems give rise to linear differential equations with coefficients which are variable with time. If average time values are adopted over successive small intervals, the complete motion can be pictured as the limit of a succession of disturbed steady motions, each associated with an appropriate set of eigenvalues.

There are, of course, many other interesting aspects of aerodynamics and structural engineering that are related to eigenvalues, but they are beyond the scope of a short survey. To close this paper, I can therefore only regret both its omissions and its treatment, which has inevitably laid greater emphasis on physical questions than on actual numerical analysis.

References

[1] W. P. Jones, Wing fuselage flutter of large aeroplanes, Rep. & Memoranda Brit. Aeronaut. Council 2656 (1947).

[2] J. Williams, Some developments of expansion methods for solving the flutter equations, Aeronaut. Quart. **2,** 209–225 (1950).

[3] R. A. Frazer, Bi-variate partial fractions and their application to flutter and stability problems, Proc. Roy. Soc. [A] **185,** 465–484 (1946).

[4] N. C. Lambourne and D. Weston, An experimental investigation of the effect of localized masses on the flutter of a model wing, Rep. & Memoranda Brit. Aeronaut. Council 2533 (1944).

[5] W. J. Duncan, A. R. Collar, and H. M. Lyon, Oscillation of elastic blades and wings in an airstream, Rep. & Memoranda Brit. Aeronaut. Council 1716 (1936).

[6] W. H. Ingram, The modal oscillations of discrete dynamical systems, Phil. Mag. [7] **38,** 51–64 (1947).

[7] W. J. Duncan, Mechanical admittances and their application to oscillation problems, Rep. & Memoranda Brit. Aeronaut. Council 2000, monograph (1947).

[8] J. P. Den Hartog, Mechanical vibrations, p. 171 (McGraw-Hill Co., New York and London, 1940).

[9] Th. Theodorsen and I. E. Garrick, Mechanism of flutter. A theoretical and experimental investigation of the flutter problem, Nat. Advisory Comm. Aeronaut. Rep. **685** (1940).

[10] J. Williams, Methods of predicting flexure-torsion flutter of cantilever wings, Rep. & Memoranda Brit. Aeronaut. Council 1990 (1943).

[11] H. Reissner, Z. Bauwesen **49,** 477 (1899).

[12] R. A. Frazer, Natural frequencies and elastic stiffnesses of bridge box structures, Rep. Aerodynamics Div., Nat. Phys. Lab., Aero **195,** 1–73 (1950).

[13] A. G. Pugsley, A flexibility-coefficient approach to suspension bridge theory, J. Instit. Civ. Eng. **32,** 226 (1949).

[14] G. I. Taylor, Stability of a viscous liquid contained between two rotating cylinders, Phil. Trans. Roy. Soc. [A] **223,** 289–343 (1923).

[15] W. P. Jones, The problem of stability of flow between rotating cylinders treated by collocation, Paper No. 3215 of the British Aeronaut. Research Council, unpublished (1937).

[16] Chi-Teh Wang and R. F. Brodsky, Application of Galerkin's method to compressible fluid-flow problems, J. Applied Phys. [12] **20,** 1255–1256 (1949).

[17] Chi-Teh Wang and R. F. Brodsky, Approximate solution of compressible fluid-flow problems by Galerkin's method, J. Aeronaut. Sci. **17,** 660–666 (1950).

[18] R. A. Frazer, On the use of polynomial approximations in hydrodynamics, Paper No. 11,672 of the British Aeronaut. Research Council, unpublished (1948).

[19] R. V. Southwell and H. B. Squire, A modification of Oseen's approximate equation for the motion in two dimensions of a viscous incompressible fluid, Phil. Trans. Roy. [A] **232,** 27–64 (1932).

9. Inclusion Theorems for Eigenvalues

H. Wielandt[1]

The Problem

Though there exists a large number of inclusion theorems for the eigenvalues of matrices, little effort seems to have been made to develop what we might call an *inclusion theory,* that is to say a systematic theory giving a full account of all possible inclusion theorems that may be based on given assumptions. The following remarks are intended to give some contributions toward such a theory.

Of course, we cannot expect to succeed in giving a unified theory covering all known inclusion theorems. There are too many different types of premises involved. We therefore shall confine our attention to one definite type of inclusion theorems, which may be considered to be the most natural one. It is that type we meet in principle every time we have solved an eigenvalue problem $Ax=\lambda x$ approximately.

Suppose we have obtained some vector x and some scalar λ_0 satisfying $Ax\approx\lambda_0 x$. Then the question arises: What do we know about the true eigenvalues of A? We may expect to find some eigenvalue λ of A to lie within some neighborhood of λ_0, the extent of this neighborhood depending on the degree of accuracy of the above approximation. This degree of accuracy should be measured in some way or another, if we want to obtain numerical inclusion theorems. There are several ways of doing this. But since we want our theory to be as general as possible, we prefer to avoid a decision in favor of one of these possibilities. So we state our problem in the following, more general form:

Let us be *given two vectors x,y of n complex components.* (For the sake of simplicity we suppose the length of x to be 1: $||x||^2=x^*x=1$, and y to be linearly independent of x). Then we ask: *What restrictions are imposed upon the eigenvalues of an n-rowed matrix A by postulating $Ax=y$?*

The first answer to this question is disappointing. For it is easy to prove that there is no restriction at all. Even if the given vectors x,y are proportional with an extremely high degree of accuracy, we may choose n arbitrary numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ and still find a matrix A possessing these values as eigenvalues and carrying x into y. So if we want to get any information about the eigenvalues of A, we have to impose on A some additional condition. Here arbitrariness is inevitable. Let us take as starting point that class of matrices whose eigenvalue theory is best known.

Hermitian Matrices

Let A be a Hermitian matrix. In this case there are several well-known inclusion theorems of the type considered. For instance, if we let the real number $x^*y=\mu$, then there is at least one eigenvalue of A to the left of μ: $\lambda_1\leq\mu$, and another to the right: $\lambda_2\geq\mu$. Further, if we define a positive number r by $r^2=||y||^2-\mu^2$, then at least one eigenvalue satisfies the inequality $|\lambda_3-\mu|\leq r$, first given by D. N. Weinstein in the case of differential equations.

Now let us attack the problem of determining all possible inclusion theorems for Hermitian matrices First of all we have to make the problem precise. This may be done by introducing two concepts, that of compatible spectra and that of inclusion sets.

We shall denote by the term *compatible spectrum* the set of the eigenvalues of every Hermitian matrix A that carries the given vector x into the given vector y. So $\lambda_1, \ldots, \lambda_n$ are a compatible spectrum, if there exists some matrix A with the properties

$$A=A^*, \quad Ax=y, \quad |\lambda I-A|=\prod_{k=1}^{n}(\lambda-\lambda_k).$$

[1] University of Tübingen, Germany

Using this definition, we may say that each of the three theorems quoted determines a certain subset of the λ-axis, which contains at least one point out of every compatible spectrum. Such a subset shall be called an *inclusion set.* (Of course both of these concepts depend on the vectors x and y. But since we shall keep these vectors fixed throughout our investigation, we shall not mention this dependence further.) Now we are able to make our problem precise: *Given x and y, to determine all inclusion sets.*

It is perhaps surprising that this problem may be settled completely. In fact, we can even go a step further and determine all compatible spectra. The solution of this problem takes its simplest form in geometrical terms. Let C denote the circle with center μ and radius r defined above. C is determined by x and y. If we project the λ-axis stereographically upon this circle, then to every spectrum $\lambda_1, \ldots, \lambda_n$ corresponds a group of n not necessarily distinct points of C. Now the following theorem holds.

A spectrum is compatible if and only if the convex closure of the corresponding point set of C contains the center μ of the circle. In other words, incompatible spectra are those that may be included in a circular segment of angle less than π. From this we can easily derive all inclusion sets:

A given subset M of C is an inclusion set if and only if it contains a semicircle. So the semicircles are the minimal inclusion sets, in the sense that we cannot omit a single point without destroying the inclusion property.

From the inclusion sets on the λ-circle C we obtain the inclusion sets on the λ-axis by simply projecting back.

Thus all possible inclusion theorems of the type considered are found and now it is an easy task to choose for every purpose and every supplementary knowledge available the most appropriate inclusion theorem.

For instance, if we are searching for the best upper bound β for the smallest eigenvalue and β is to depend on x and y only, choose the minimal inclusion set for which the upper bound is as low as possible. This becomes the half-axis $\lambda \leq \mu$ hence $\beta = \mu$. Or if we are interested in finding an approximate eigenvalue, for which we can guarantee a minimal absolute error, then we have to take the center of the inclusion interval of smallest length; this is Weinstein's interval; so the approximate eigenvalue with smallest absolute error is μ.

The Hermitian case being settled, we look for generalizations.

Normal Matrices

In the Hermitian case the proofs are based on the fact that Hermitian matrices are characterized by $A = U^{-1}DU$, D real diagonal, U unitary.

Since the reality condition is not very important in the proof, the easiest way for generalization is to drop this condition. It is well known that the remaining conditions characterize the *normal* matrices. If we generalize the concepts of compatibility and inclusion sets by admitting normal matrices instead of Hermitian matrices, and by considering complex spectra and inclusion sets, then the proofs run exactly as before, and we arrive at the following result:

Let the complex number μ and the positive number r be defined by

$$\mu = x^*y, \quad r^2 = \|y\|^2 - |\mu|^2.$$

Let the complex plane be projected stereographically upon the sphere S with center μ and radius r. Then a given spectrum is compatible if and only if the convex closure of the corresponding point set on the sphere contains the center μ of the sphere. In other words, exactly those spectra are incompatible that may be included in a spherical segment less than a hemisphere. Now again, it is easy to determine all inclusion sets: *A given subset M of S is an inclusion set if, and only if, it contains a hemisphere.* The hemispheres are the minimal inclusion sets yielding the optimal inclusion theorems. So the normal case is settled in a satisfactory way, and we may turn to further generalization.

Normalizable Matrices

Let us drop the condition that U be unitary. We obtain the class of all matrices that are conjugate to diagonal matrices. They may be described in other words as the matrices possessing n linearly independent eigenvectors or as the matrices possessing only linear elementary divisors. Let me refer to these matrices as the *normalizable* matrices. This class of matrices is a very extensive one, although it still does not contain all of them. In fact, it is too large in order to yield inclusion theorems valid throughout this class.

Here a new idea enters. We may hope to obtain inclusion theorems again, if we define some measure for the deviation of a given normalizable matrix from normality and then restrict the class of matrices to be considered by limiting the deviation. There are many ways of measuring deviation from normality, but it is not quite easy to find a definition suitable to eigenvalue inclusion theory. We proceed as follows. We define real angles between any two vectors a,b of the complex linear space in which the eigenvectors are to be found in the following way:

$$\cos \sphericalangle(a,b)=\frac{|a^*b|}{\|a\|\|b\|}, \quad 0\leq \sphericalangle(a,b)\leq\frac{\pi}{2}.$$

Then we consider all the matrices U transforming A into diagonal form. If one of these matrices preserves all right angles, then it is a scalar multiple of a unitary matrix; in this case A is normal. This suggests to measure the deviation of A from normality in the general case by measuring the deformation of right angles, which is necessary for transforming A into diagonal form. To carry this out, we define for each nonsingular matrix U a *deformation angle* $\delta_U=\min \sphericalangle(Ux,Uy)$, $\left(\sphericalangle(x,y)=\frac{\pi}{2}\right)$, and for each normalizable A we define a *normality angle* $\nu_A=\max \delta_U$, $(U^{-1}AU=\text{diagonal})$. So we have $0<\nu_A\leq\frac{\pi}{2}$, where the equality sign holds for normal matrices only. A small value of ν_A means a large deviation of A from normality.

Now we may generalize the former concept of compatibility to what we shall call ν-compatibility, where ν denotes some fixed angle in the interval $0<\nu\leq\pi/2$. A set of n complex numbers $\lambda_1, \ldots, \lambda_n$ shall be called a *ν-compatible spectrum*, if there exists a normalizable matrix A with the properties

$$Ax=y, \quad |\lambda I-A|=\prod_{k=1}^{n}(\lambda-\lambda_k), \quad \nu_A\geq\nu.$$

We propose to determine all ν-compatible spectra. The solution of this problem is described most conveniently, using the same number sphere S determined by x and y that we introduced in the normal case. For the sake of convenience, let us define the spherical radius $\rho(\lambda_1, \ldots, \lambda_n)$ of a spectrum to be the spherical radius of the smallest segment of S containing all the λ_ν. Then we may state the following theorem.

Main theorem: $\lambda_1, \ldots, \lambda_n$ *are a ν-compatible spectrum if and only if* $\rho(\lambda_1, \ldots, \lambda_n)\geq\nu$.

In other words, suppose we are given vectors x,y and any spectrum $\lambda_1, \ldots, \lambda_n$ possessing any spherical radius ρ upon the number sphere determined by x and y. Then there exists a matrix A with normality angle $\nu_A=\rho$, which possesses the spectrum $\lambda_1, \ldots, \lambda_n$ and carries x into y; but there is no such matrix with normality angle $\nu_A>\rho$.

The ν-compatible spectra being known, there is no difficulty in determining the corresponding inclusion sets. *The minimal inclusion sets are the spherical segments with spherical radius $\pi-\nu$; every other inclusion set contains a minimal one.* So all possible inclusion theorems are known, and it would be easy to generalize all the well-known special inclusion theorems holding for real symmetric matrices to the case of normalizable matrices.

Generalizations

After the normalizable case, we ought to consider the most general case, imposing on A no other restriction than a certain limitation of the deviation from normality. This case is rather more difficult than the foregoing, owing to the fact that now elementary divisors of higher degree are allowed to occur. As this investigation has not yet been finished, I should like to confine myself to indicating the method employed. It is essentially a geometrical one.

The main theorem stated before implies a remarkable geometrical fact. We remember that the normality angle ν was defined to be an angle in the n-dimensional space, in which the eigenvectors are to be found. In the theorem, the same angle appears in a certain three-dimensional space, namely, the interior of a certain complex number sphere, upon which the eigenvalues are to be found. To put it briefly, one and the same angle plays a fundamental role in the eigenvector space and in the eigenvalue space. This surprising result, found by calculation, caused me to search for immediate geometrical connections between the eigenvector space and the eigenvalue space. Such relations exist indeed. The link connecting the eigenvector space V and the eigenvalue space N is the three-dimensional set F of those points, whose rectangular coordinates may be expressed by $\xi = \text{Re } x^*Ax$, $\eta = \text{Im } x^*Ax$, $\zeta = x^*A^*Ax$, where x runs through all vectors of length 1 in V. F is a convex set containing on its boundary the n points P_ν with coordinates $\xi = \text{Re } \lambda_\nu$, $\eta = \text{Im } \lambda_\nu$, $\zeta = |\lambda_\nu|^2$ corresponding to the spectrum of A. F is the convex closure of these n points if and only if A is normal. F lies in the interior of the paraboloid P with the equation $\zeta = \xi^2 + \eta^2$. P can be mapped upon the number sphere S by an appropriate collineation In this way a mapping of the eigenvector space V into the eigenvalue space N is defined. The geometrica relations between these two spaces may be expressed most adequately in terms of hyperbolic geometry At the present time, they provide the most powerful tool for research in this field.

I should like to finish by indicating two further generalizations of the problem we dealt with. The first consists in assuming more than one vector and its image to be given, say $x_1, \ldots, x_s$ and $Ax_1, \ldots, Ax_s$. What then can be said about the spectrum of A? Here even the case of Hermitian matrices is highly complicated and up to the present time not perfectly cleared up. However, in the important special case, that some iterates of one single vector are given, say x_0, $Ax_0 = x_1, \ldots, Ax_{s-1} = x_s$ $(A = A^*)$, a satisfactory survey of all inclusion sets has been obtained. For example, I mention that for every given real number ξ the shortest inclusion intervals with left or right end ξ, say $\eta_1 \leq \lambda < \xi$ and $\xi < \lambda \leq \eta_2$, can be calculated by solving the following algebraic equation of degree s for the two roots η_1, η_2 nearest to ξ:

$$\begin{vmatrix} m_0 & m_1 & \ldots & m_s & 1 \\ m_1 & m_2 & \ldots & m_{s+1} & \eta \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ m_s & m_{s+1} & \ldots & m_{2s} & \eta^s \\ 1 & \xi & \ldots & \xi^s & 0 \end{vmatrix} = 0.$$

The matrix elements are defined by $m_k = x_i^* \, x_j \quad (i+j=k)$.

The second extension I want to mention consists in considering inclusion theorems not only for *eigenvalues*, but for eigen*vectors* too. Because this problem goes beyond the limits of the present talk, I shall confine myself to mention the rather surprising fact that it can be solved in a perfectly satisfactory way for those classes of matrices for which a full eigenvalue inclusion theory has been developed.

There is no difficulty in extending the theory to the case of operators in Hilbert space.

10. On General Computation Methods For Eigenvalues and Eigenfunctions[1]

Gaetano Fichera[2]

Methods for computation of eigenvalues of symmetric or Hermitian quadratic operators in Hilbert spaces have now been so fully developed, that eigenvalue problems arising in the applications, related to self-adjoint linear equations, can be considered—at least from a theoretical standpoint—as solved. However, in some cases computation of eigenvalues of nonsymmetric operators is needed.

We want to sketch here two procedures that have been applied at the Italian Institute for the Applications of Calculus in connection with quite general linear eigenvalue problems. The first of these can also successfully be applied even in self-adjoint problems, in which classical procedures are impractical because of the difficulty of getting complete sets of functions satisfying certain boundary conditions.

General Least Squares Method

Let ***H*** be a separable Hilbert space, linear with respect to the complex field. If u is an element of ***H***, $\bar{u}$ denotes the complex conjugate of u. Let us denote in the usual way the scalar product of two elements u and v of ***H*** by (u,v). It is well known that the scalar product has the following properties:

(a) $(u,v)=(v,u)$,

(b) $(u,\alpha v_1+\beta v_2)=\alpha(u,v_1)+\beta(u,v_2)$, ($\alpha$ and β complex numbers),

(c) $(\bar{u},\bar{v})=(\overline{u,v})$,

(d) $\|u\|^2=(u,\bar{u})$ is real and positive when u does not coincide with the origin ω of ***H***.

Let $L[u]$ and $M[u]$ be two operators that map ***H*** or part of it, respectively, in two linear subspaces of ***H***. Then

$$L[\alpha u_1+\beta u_2]=\alpha L[u_1]+\beta L[u_2]$$

$$M[\alpha u_1+\beta u_2]=\alpha M[u_1]+\beta M[u_2].$$

In addition we suppose

$$\lim_{\|u\|\to 0}\|L[u]\|=0, \qquad \lim_{\|u\|\to 0}\|M[u]\|=0.$$

We start with the following eigenvalue problem

$$L[u]+\lambda M[u]=\omega \qquad \|u\|=1. \tag{1}$$

Let us consider the following real-valued functional

$$P(u,\lambda)=\|L[u]+\lambda M[u]\|^2.$$

It is evident that $P(u,\lambda)$ vanishes on the unit-sphere $\|u\|=1$ when and only when an eigenvalue of the problem (1) exists. Let $\{v_k\}$ then be a complete system of points of ***H***.

Generally the set of the eigenvalues for problem (1) is a set of points of the complex plane without finite limit point. We shall assume this hypothesis for problem (1) and denote by $\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots$, the sequence of the eigenvalues ordered in such way that $|\lambda_n|\leq|\lambda_{n+1}|$ and if $|\lambda_n|=|\lambda_{n+1}|$, then princ. arg $\lambda_n<$ princ. arg λ_{n+1}.

Let us set

$$u^{(n)}=\sum_{k=1}^{n}c_k v_k, \tag{2}$$

$$P_n(c_1,c_2,\ldots,c_n,\lambda)=P(u_n,\lambda)=\sum_{h,k}^{1,n}A_{hk}c_h\bar{c}_k,$$

[1] The methods discussed in this paper were given by Prof. M. Picone and applied at the Italian Institute for the Applications of Calculus.

[2] University of Trieste and Italian Institute for the Applications of Calculus.

$$A_{hk}=(L[v_h],\overline{L[v_k]})+\lambda(M[v_h],\overline{L[v_k]})+\bar{\lambda}(\overline{M[v_k]},L[v_h])+\lambda\bar{\lambda}(M[v_h],\overline{M[v_k]}),$$

$$Q_n(c_1,c_2,\ldots,c_n)=||u^{(n)}||^2=\sum_{k=1}^{n}|c_k|^2.$$

P_n and Q_n are two Hermitian quadratic forms in the variables $c_1, c_2, \ldots, c_n$.

We want to get the minimum of P_n under the condition $Q_n=1$. We are thus led according to the theory of Hermitian forms to the following linear algebraic system in the unknowns $c_1, c_2, \ldots, c_n$:

$$\sum_{k=1}^{n} A_{hk}\bar{c}_k-\mu\bar{c}_h=0. \tag{3}$$

Let $D(\lambda,\bar{\lambda},\mu)$ be the determinant of system (3). We get the equation

$$D(\lambda,\bar{\lambda},\mu)=0. \tag{4}$$

For every fixed λ, equation (4) has, as is well known, n real roots.

If (3) and (4) are satisfied, then $P_n-\mu Q_n=0$. From the condition $Q_n=1$, it follows $P_n=\mu$. Let $\lambda_1^{(n)}$, $\lambda_2^{(n)}$, $\ldots$, $\lambda_{\nu_n}^{(n)}$ be the minimizing points for the functions $\mu(\lambda)$ implicitly defined by (4). We suppose that these values are ordered in the same way as the sequence $\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots$. The values so obtained are taken as nth approximation of the first ν_n eigenvalues of problem (1).

Of course, a theoretical justification of this procedure is needed and the method at the actual state is quite empirical. However, the numerical experiences performed in the Italian Institute in many particular problems gave always very satisfactory results not only for the determination of the eigenvalues but also for the computation of the corresponding eigenfunctions.

We want to report here the numerical results obtained in a problem, in which other procedures are applicable too, in order to compare the results of all of them.

The problem, which is connected with an engineering problem, is the following:[3]

$$\frac{d^2}{dx^2}\left[(1-\theta x)^3\frac{d^2u}{dx^2}\right]-\lambda(1-\theta x)u=0, \qquad \int_0^1 u^2dx=1, \qquad u(0)=u(1)=u''(0)=u''(1)=0. \tag{5}$$

θ is a numerical nonnegative constant less than 1. In the case $\theta=0$, the eigenvalues λ_n and the corresponding eigenfunction $u_n(x)$ are known:

$$\lambda_n=n^4\pi^4, \qquad u_n(x)=\sqrt{2}\sin n\pi x.$$

The first approximation gives the following approximation for the first eigenvalue

$$\lambda_1\simeq\frac{3024}{31}\left(1-\theta-\frac{\theta^2}{42}\right).$$

In the case $\theta=0$, we get $\lambda_1\simeq97.548$, while the exact value is $\lambda_1=\pi^4=97.4091$. The following table shows the numerical results obtained in the case $\theta=0$ for the first three eigenvalues, compared with their exact values.

	First approximation	Second approximation	Third approximation	Exact value
λ_1	97.548	97.548	97.4091	97.4091
λ_2	------------	1584.004	1584.004	1558.545
λ_3	------------	------------	7890.136	7890.13

[3] T. Viola, Calcolo approssimato di autovalori, Rend. di Mat. e delle sue appl. S. V. II, (1941).

The following table compares the values obtained by the method in second and third approximation for the first eigenfunction with the exact values:

	$\frac{1}{\sqrt{2}}u_1(x)$		$\sin \pi x$
	Second approximation	Third approximation	
0	0	0	0
0. 1	0. 308 190	0. 309 003	0. 309 017
0. 2	0. 583 080	0. 587 644	0. 587 785
0. 3	0. 798 279	0. 808 599	0. 809 017
0. 4	0. 934 938	0. 950 353	0. 951 057
0. 5	0. 981 748	0. 999 160	1
0. 6	0. 934 938	0. 950 353	0. 951 057
0. 7	0. 798 279	0. 808 599	0. 809 017
0. 8	0. 583 080	0. 587 644	0. 587 785
0. 9	0. 308 190	0. 309 003	0. 309 017
1	0	0	0

For $\theta=0.5$ we have the following table of the values obtained in four approximations for the first three eigenvalues:

	First approximation	Second approximation	Third approximation	Fourth approximation
λ_1	48. 194	48. 720	50. 890	50. 8129
λ_2	------------	852. 95	838.	836. 9412
λ_3	------------	------------	------------	4256. 878

Tricomi [4] gives for λ_1 the lower bound 47.9215 and the upper bound 57.1553. We shall compare the obtained numerical results with those obtained by using the method of the next section.

Method Founded on Cauchy-Lipschitz Integration Procedure

The present method is in connection with eigenvalue problems for normal systems of ordinary linear differential equations. We can suppose the system to be of the first order, and we write it in the following vectorial form

$$\frac{dx}{dt}=H(t,\lambda)*x(t). \tag{6}$$

$x(t)$ is a vector function of the real variable t, in an interval (α,β) having p real or complex components $x_h(t)$ $(h=1,2,\ldots,p)$. $H(t,\lambda)$ is a square matrix of order p. Its elements $H_{hk}(t,\lambda)$ depend continuously upon the real variable t and the complex variable λ. $H(t,\lambda)*x(t)$ denotes the p-vector having as hth component $\sum_{k=1}^{p} H_{hk}(t,\lambda)x_k(t)$.

Let $L_1[x]$, $L_2[x]$, . . ., $L_p(x)$ be p functionals depending linearly upon x, that is to say,

$$L_i[ax+by]=aL_i[x]+bL_i[y]$$

for every pair of complex constants a and b.

[4] F. Tricomi, Sulle vibrazionai trasversali di aste, specialmente di bielle di sezione variabile, Ricerche di Ingegneria [2] **IV**, 47 (Marzo-Aprile 1936).

We consider the following eigenvalue problem

$$\frac{dx}{dt}=H(t,\lambda)*x(t)$$

$$L_i[x]=0 \quad (i=1,2,\ldots,p), \quad \int_\alpha^\beta |x(t)|^2dt=1. \tag{7}$$

Let $x^{(k)}(t,\lambda)$ be the integral of (6) determined by the initial condition

$$x_h^{(k)}(\alpha)=\delta_h^k. \tag{8}$$

The general integral of (6) is

$$x(t,\lambda)=\sum_{k=1}^{p} c_k x^{(k)}(t,\lambda).$$

A necessary and sufficient condition for the existence of eigenvalues of problem (7) is the existence of roots of the equation

$$\Delta(\lambda)=||L_i[x^{(k)}(t,\lambda)]||=0.$$

Let $x^{(k,n)}(t,\lambda)$ be the vector obtained by applying the Cauchy-Lipschitz approximation method in the nth step with the initial condition (8).

Let us set

$$x^{(n)}(t,\lambda)=\sum_{k=1}^{p} c_k^{(n)}x^{(k,n)}(t,\lambda), \text{ and } \Delta_n(\lambda)=||L_i[x^{(k,n)}(t,\lambda)]||=0.$$

If $L_i[x](i=1,2,\ldots,p)$ is supposed to be a continuous functional of the first order of x, that is to say if $\lim|L[x]|=0$, if $(|x|+|x'|)\to 0$, and if $H_{hk}(t,\lambda)$ has continuous first derivatives with respect to t and is an holomorphic function of λ in the whole complex plane, then it has been shown[5] that every root of $\Delta(\lambda)$ is approximated by a root of $\Delta_n(\lambda)$.

The method described was applied to the problem (5). Putting $n=10$, that is to say, by applying the Cauchy-Lipschitz procedure when the interval (α,β) is divided in 10 parts, it furnished the following values for the first three eigenvalues in the case $\theta=0.5$: $\lambda_1=50.9$, $\lambda_2=837.4$, $\lambda_3=4212.2$. The following table compares the approximate value for the first eigenfunction of the problem (5), obtained by least squares method and Cauchy-Lipschitz procedure in the case $\theta=0.5$:

x	$u_1(x)$	
	Least squares	Cauchy-Lipschitz
0	0	0
0. 1	0. 099 17	0. 099 17
0. 2	0. 192 94	0. 192 97
0. 3	0. 274 86	0. 274 91
0. 4	0. 337 70	0. 337 67
0. 5	0. 373 86	0. 373 58
0. 6	0. 376 16	0. 375 47
0. 7	0. 338 95	0. 337 87
0. 8	0. 259 92	0. 258 69
0. 9	0. 142 62	0. 141 63
1	0	0

[5] T. Viola, Dimostrazione della convergenza di un procedimento di M. Picone per il calcolo degli autovalori, Rend Accademia Naz Lincei XXIX (1939).

11. Variational Methods for the Approximation and Exact Computation of Eigenvalues[1]

Alexander Weinstein[2]

11.1. Introduction

The variational methods for the approximation of eigenvalues have always been intimately connected with the development of the theory of membranes and plates. In fact, the variational definition of eigenvalues so often used today goes back to a classical paper of H. Weber on the vibration of membranes which was published as the first paper in the first volume of the Mathematische Annalen. The methods of Rayleigh-Ritz were first developed for plates. In the decade 1910–20 the minimax theory and the asymptotic laws for the distribution of eigenvalues were first developed for membranes and plates [1, 2]. More recently, the same problems led to methods for the variational determination of upper and lower bounds for eigenvalues which resulted in a unified theory of eigenvalues of plates and membranes [3, 4].

Although it would therefore be justified to present the theory of the approximation of eigenvalues in terms of differential operators, we shall use mostly the theory of integral operators which are the inverse of the differential operators. In other words, we shall deal with completely continuous operators in a Hilbert space. In the present paper we will consider only methods that lead to arbitrarily precise approximations of eigenvalues, and will not be concerned with rough inequalities.

11.2. Operators in a Hilbert Space and Its Subspaces

Let $\boldsymbol{H}$ be a real Hilbert space and let L be a positive definite completely continuous symmetric operator. We shall use the standard notations of the theory of Hilbert space. The eigenvalue equation

$$Lu=\lambda u \tag{1}$$

possesses an infinite number of positive eigenvalues

$$\lambda_1^{(0)}\geq\lambda_2^{(0)}\geq\ .\ .\ .;\quad \lim_{n\to\infty}\lambda_n^{(0)}=0 \tag{2}$$

with the corresponding normalized eigenvectors

$$u_1^{(0)},\quad u_2^{(0)},\quad .\ .\ .\ . \tag{3}$$

It is well known that these eigenvalues can be obtained by a chain of maximum problems. For instance,

$$\lambda_1^{(0)}=\max\frac{(u,Lu)}{(u,u)}\text{ for }u\text{ in}\boldsymbol{H}. \tag{4}$$

The second eigenvalue $\lambda_2^{(0)}$ is the maximum of the same expression for u orthogonal to $u_1^{(0)}$, and so forth.

Let $\boldsymbol{Q}$ be a closed linear subspace of $\boldsymbol{H}$, and

$$\boldsymbol{P}=\boldsymbol{H}\ominus\boldsymbol{Q}. \tag{5}$$

Consider the corresponding chain of maximum problems for the expression $(v,Lv)/(v,v)$ where v is restricted to $\boldsymbol{Q}$. The maximizing vectors again satisfy an eigenvalue equation of the type

$$Lv-\lambda v=p, \tag{6}$$

[1] This work was performed under the sponsorship of the Office of Naval Research.

[2] Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, College Park, Maryland

where p is some element of $\boldsymbol{P}$. In order to put (6) in the same form as (1), we take the projections of both sides into the space $\boldsymbol{Q}$. The projection of Lv is itself a completely continuous positive operator in $\boldsymbol{Q}$, which we call $L'v$. In this way, we obtain in place of (6) the eigenvalue equation

$$L'v-\lambda v=0 \tag{7}$$

which admits the eigenvalues

$$\lambda_1'\geq\lambda_2'\geq\ .\ .\ .;\quad \lambda_n'\to 0 \tag{8}$$

with the corresponding eigenvectors

$$u_1',\quad u_2',\quad .\ .\ .\ . \tag{9}$$

The minimax theory gives the fundamental inequalities

$$\lambda_n^{(0)}\geq\lambda_n';\quad n=1,\quad 2,\quad .\ .\ .\ . \tag{10}$$

Let us note that these well-known inequalities include the Rayleigh-Ritz method. In fact, if $\boldsymbol{Q}$ is a space of a finite number of dimensions, the operator L' is a matrix, so that the eigenvalues $\lambda_n^{(0)}$ of L' can be considered for the purpose of the present paper as known. In this way, lower bounds for the $\lambda_n^{(0)}$ are given by (10). (It should be noted that the eigenvalues of a differential operator are the inverses of the eigenvalues of the corresponding L, so that the signs of all inequalities must be reversed.)

Generally speaking, the following alternatives may be encountered. Either the eigenvalues and eigenvectors of L or those of L' are known. In the first case, when the $\lambda_n^{(0)}$ are known, (10) gives upper bounds for the eigenvalues λ_n' of L'. In the second case, (10) gives lower bounds for the eigenvalues $\lambda_n^{(0)}$ of L. Let us give an example for the first case. Let G be the Green's function for a square satisfying the equation $\Delta\Delta G=0$ and the boundary conditions $G=\Delta G=0$, and define L as the integral operator whose kernel is G. Then the $\lambda_n^{(0)}$ are explicitly known, since the problem is then essentially that of a vibrating membrane. For L' we take as the kernel the Green's function satisfying the same equation $\Delta\Delta G'=0$ and the boundary conditions $G'=\partial G'/\partial n=0$. This last problem is the problem of the vibrating clamped square plate. The subspace $\boldsymbol{P}$ turns out to be the space of harmonic functions of integrable square.

For the second case, an example where λ_n' is known is given by the problem of the vibrations of a circular clamped plate. The operator L in this case refers to the vibration of a circumscribed square plate. In this case, the subspace $\boldsymbol{P}$ is the space of functions harmonic within the circle and arbitrary in the remainder of the square.

11.3. The Intermediate Problems

We now introduce the concept of intermediate problems, which will lead to an essential improvement of the bounds given by the inequality (10). We shall restrict ourselves to the consideration of upper bounds. This will be done not only to avoid repetition, but also because some of the new results given in this paper have only been derived for upper bounds. In other words, we consider the first case of the preceding section, where the eigenvalues $\lambda_n^{(0)}$ and the eigenfunctions $u_n^{(0)}$ of L are known. Let $(p_1, p_2,\ .\ .\ .\ )$ be a basis for the subspace $\boldsymbol{P}$, that is, a complete set of linearly independent (but not necessarily orthogonal) elements of $\boldsymbol{P}$. There is an infinite choice of such bases, but for the time being we will not specify any particular one.

The intermediate problem of index m is defined in the following way. Introduce the space

$$\boldsymbol{H}_m=\boldsymbol{H}\ominus\{p_1,\ .\ .\ .\ ,\ p_m\}. \tag{11}$$

Consider the chain of maximum problems for the expression

$$(v,Lv)/(v,v) \tag{12}$$

with v restricted to lie in $\boldsymbol{H}_m$. This leads to an eigenvalue problem of the form (6), where p must in the present case be a vector in the finite space $\{p_1,\ .\ .\ .\ ,\ p_m\}$. In this way, we get the equation

$$Lu_n^{(m)}-\lambda_n^{(m)}u_n^{(m)}=a_n^{m,1}p_1+\ .\ .\ .\ -a_n^{m,m}p_m. \tag{13}$$

where the constants $a_n^{m,i}$ are unknown. Again taking the projection into $\boldsymbol{H}_m$ ansd calling $L^{(m)}$ the projection of L into $\boldsymbol{H}_m$, this can be written, similar to (7), as

$$L^{(m)}u-\lambda u=0. \tag{14}$$

This equation yields the eigenvectors $u_n^{(m)}$ and the eigenvalues

$$\lambda_1^{(m)}\geq\lambda_2^{(m)}\geq\ .\ .\ .;\quad \lim_{n\to\infty}\lambda_n^{(m)}=0. \tag{15}$$

Since $\boldsymbol{H}_m$ contains the subspace $\boldsymbol{Q}$ and is contained in $\boldsymbol{H}_{m-1}$ and in $\boldsymbol{H}$, we have by the minimax theory, as in (10),

$$\lambda_n^{(0)}\geq\lambda_n^{(1)}\geq\lambda_n^{(2)}\geq\ .\ .\ .\ \geq\lambda_n';\quad n=1,2,\ .\ .\ .\ . \tag{16}$$

It has been shown that for each fixed n

$$\lim_{n\to\infty}\lambda_n^{(m)}=\lambda_n'. \tag{17}$$

It would perhaps have been more appropriate to use the notation $\boldsymbol{H}_0$ for $\boldsymbol{H}$ and $\boldsymbol{H}_\infty$ for $\boldsymbol{Q}$ and to write in place of L and L' the symbols $L^{(0)}$ and $L^{(\infty)}$, but this would be too cumbersome.

The intermediate problem was introduced by Weinstein [3]. The question of convergence was investigated jointly by Aronszajn and Weinstein [4].

The fundamental fact about the intermediate problems is the following: The eigenvalues and eigenfunctions of the operator $L^{(m)}$ can be explicitly computed in terms of the corresponding quantities of L, which are assumed to be known. The problem for L could therefore be called the base problem. The problem for $L'(=L^{(\infty)})$, cannot be explicitly solved, but appears as a limiting case of explicitly solvable intermediate problems.

The discussion and solution of the intermediate problems is obviously a major part of the theory. It was given for a special sequence $(p_1,p_2,\ .\ .\ .)$ by Weinstein [3] in the case of the vibrating plate, and later substantially improved by Aronszajn [5, 6] for general operators in Hilbert space. Aronszajn rightly emphasized the fact that the solution of the intermediate problems can be given by the use of an arbitrary basis $(p_1,p_2,\ .\ .\ .)$. We shall refrain from the formulation of these general results, inasmuch as adequate presentation of this part of the theory has been given in recent papers mentioned above.

The method of upper bounds was originally introduced by Weinstein as the counterpart of the Rayleigh-Ritz method for lower bounds. Later, Aronszajn introduced as its actual counterpart a generalized Rayleigh-Ritz method in which the space $\boldsymbol{Q}$ defined in section 11.2 may have an infinite number of dimensions. He showed that the generalized Rayleigh-Ritz method can be discussed in essentially the same fashion as the method of upper bounds [5, 6, 7]. The basic idea underlying all these new developments is the consideration of spaces that differ only by a finite-dimensional space.

The main purpose of the following paragraphs will be to bring some new and partly unpublished results going beyond the question of solution of the intermediate problems and the convergence of the approximations. These results are obtained by a special selection of the basis $(p_1, p_2\ .\ .\ .\ )$.

11.4. Weinberger's Estimate of the Error in the Method for Upper Bounds

As long as the Rayleigh-Ritz method was the only known variational method, the error in the lower bounds was not known. By combining the methods for lower bounds (either Rayleigh-Ritz or generalized Rayleigh-Ritz as introduced by Aronszajn) with the method for upper bounds, we can get in numerical applications an approximation to the eigenvalues with a known error which can be made arbitrarily small. However, the number of steps required to obtain a prescribed precision is not known a priori, and is left to chance.

The situation would naturally be quite different if an estimate of the error were known for at least one of the variational methods. Several attempts were made for the estimation of the error in the classical Rayleigh-Ritz method but, as far as is known to the author, no such useful error estimate exists. The same is true for the generalized Rayleigh-Ritz method. However, quite recently H. F. Weinberger [8] has given an error estimate for the method of upper bounds which is based on a special selection of the basis $(p_1, p_2, \ldots)$. This estimate renders, in principle, the method for upper bounds independent of the methods for lower bounds, and immediately gives the number of steps required for a given precision.

Weinberger's result can be formulated as follows in the notation of the previous paragraphs.

Let p_n be the projection into the space $\boldsymbol{P}$ of the eigenvector $u_n^{(0)}$. Then the following inequalities hold for the intermediate problem of index m.

$$\lambda_n^{(m)} \geq \lambda_n' \geq \lambda_n^{(m)} - \lambda_{m+1}^{(0)}. \tag{18}$$

This is a uniform estimation of the error in the upper bound for the eigenvalues λ_n', which is obtained by the use of the intermediate problem of index m. The error estimate, which is equal to the known eigenvalue $\lambda_{m+1}^{(0)}$ of L, goes to zero for increasing m, uniformly in n. Of course, for each m, the lower bound for λ_n' given by (18) will be negative, and hence trivial, for sufficiently large n. But, on the other hand, for each eigenvalue λ_n' the index m of the intermediate problem can be chosen so large that the lower bound in (18) is positive.

To prove Weinberger's estimate, we again consider as in paragraph 3, the space

$$\boldsymbol{H}_m = \boldsymbol{H} \ominus \{p_1, \ldots, p_m\} \tag{19}$$

with the p_n defined as in the statement of the theorem. Furthermore, we define the space $\boldsymbol{P}_m$ by the equation

$$\boldsymbol{P}_m = \boldsymbol{P} \ominus \{p_1, \ldots, p_m\}. \tag{20}$$

Then it is obvious that

$$\boldsymbol{P}_m = \boldsymbol{H}_m \ominus \boldsymbol{Q}. \tag{21}$$

Let us call $\mu_n^{(m)}$ the eigenvalues of the projection of L into $\boldsymbol{P}_m$. Then, by a special case of the fundamental inequality due to Aronszajn [5, p. 476, corollary I'],

$$\lambda_n^{(m)} \leq \lambda_n' - \mu_1^{(m)}. \tag{22}$$

For an arbitrary choice of the basis $(p_1, p_2, \ldots)$, we have two unknowns in this inequality, namely, λ_n' and $\mu_1^{(m)}$. But for Weinberger's choice of the p_n, the space $\boldsymbol{P}_m$ is orthogonal to the eigenvectors $u_1^{(0)}, \ldots, u_m^{(0)}$. For we have, for any vector p in $\boldsymbol{P}_m$,

$$(p, u_n^{(0)}) = (p, p_n) = 0 \quad n = 1, \ldots, m. \tag{23}$$

Therefore, the classical variational definition of eigenvalues, applied to L, yields the inequality

$$\frac{(Lp,p)}{(p,p)} \leq \lambda_{m+1}^{(0)} \quad \text{for} \quad p \epsilon \boldsymbol{P}_m, \tag{24}$$

and hence

$$\mu_1^{(m)} \leq \lambda_{m+1}^{(0)}. \tag{25}$$

This, together with (22) and (16), yields the inequality (18).

The problem of the determination of error estimates is therefore reduced to the determination of the projections p_n of the eigenvectors $u_n^{(0)}$, which will be discussed elsewhere.

11.5. The Optimum Problem

Since the intermediate problems depend upon the choice of the basis $(p_1, p_2, \ldots)$, the following question arises naturally. This question will be formulated here only in the simplest case of the first intermediate problem. *What is the lowest value of the upper bound $\lambda_n^{(1)}$ for λ_n' that can be obtained by a* *suitable choice of* p_1? In this discussion n is a fixed index. Before answering this question, let us observe that the following inequality is known as the result of a separation theorem for eigenvalues [5, p. 476, theorem B]:

$$\lambda_n^{(1)} \geq \lambda_{n+1}^{(0)}. \tag{26}$$

Also, since $\lambda_n^{(1)}$ is an upper bound for λ_n', it must satisfy the inequality

$$\lambda_n^{(1)} \geq \lambda_n'. \tag{27}$$

As an answer to the question, H. F. Weinberger has shown that the vector p_1 can be so chosen that the weaker of the two inequalities (26) and (27) becomes an equality. In other words, either

$$\lambda_n^{(1)} = \lambda_{n+1}^{(0)}, \tag{28}$$

or

$$\lambda_n^{(1)} = \lambda_n'. \tag{29}$$

The proof, which involves the discussion of several cases, will be published elsewhere [9].

The solution of similar optimum problems is of importance to the computer who has no elaborate equipment at his disposal.

11.6. An Extension of the Classical Sturm-Liouville Theory

As was noted in section 11.3, the eigenvalues and eigenvectors of an intermediate problem can be exactly computed in terms of those of the base problem and vice versa. The theory of intermediate problems was originally introduced in view of applications to partial differential equations. In this field, they play an auxiliary role, linking by an infinite chain the base problem to the problem which is to be solved. However, in the theory of ordinary differential equations, the intermediate problems have an independent significance.

Some important chapters of the classical Sturm-Liouville theory deal with the relations between the eigenvalues and eigenfunctions of two problems having the same ordinary differential equation but different boundary conditions. It was recently pointed out by Weinstein [10, 11] that, for a large class of problems, the two problems in question can be interpreted as a base problem and an intermediate problem which terminates a finite chain. By the general theory, the eigenvalues and eigenfunctions of either of the two problems can be explicitly computed in terms of the other. This result yields a new proof of the classical separation theorems that correspond to the intermediate problems of index one, but goes beyond that by replacing classical inequalities by precise equations. It also yields separation theorems of higher order which, to our knowledge, were not discussed in the classical Sturm-Liouville theory.

Recently, H. F. Weinberger [12] extended the explicit solution of one differential problem in terms of another to all pairs of self-adjoint problems having the same ordinary differential operator but different boundary conditions. Naturally, in some of these cases, the separation theorems are no longer valid. In this generality, the problems cannot always be interpreted as eigenvalue problems belonging to a space and its subspace. However, it can be shown that they correspond to two spaces which are subspaces of finite index of their union. Besides its application to Sturm-Liouville theory, these new developments constitute a remarkable extension of the theory of intermediate problems.

11.7 References

[1] R. Courant, Variational methods for the solutions of problems of equilibrium and vibrations, Bul. Am. Math. Soc. **49,** 1–23 (1943).

[2] H. Weyl, Ramifications, old and new, of the eigenvalue problem, Bul. Am. Math. Soc. **56,** 115–139 (1950).

[3] A. Weinstein, Etudes des spectres des équations aux dérivées partielles, Mémorial des Sciences Mathématiques, No. 88 (Paris, 1937).

[4] N. Aronszajn and A. Weinstein, On the unified theory of eigenvalues of plates and membranes, Am. J. Math, **64,** 623–645 (1942).

[5] N. Aronszajn, Rayleigh-Ritz and A. Weinstein methods for approximation of eigenvalues, Proc. Nat. Acad. Sci. **34,** 474–480 and 594–601 (1948).

[6] N. Aronszajn, The Rayleigh-Ritz and the Weinstein methods for approximation of eigenvalues, I. Operators in a Hilbert space (Oklahoma A. and M. College, Stillwater, Okla., 1949).

[7] N. Aronszajn, Approximation methods for eigenvalues of completely continuous operators, Proc. Symposium on Spectral Theory and Differential Problems, p. 179–202 (Oklahoma A. and M. College, Stillwater, Okla., 1951).

[8] H. F. Weinberger, Error estimation in the Weinstein method for eigenvalues, Proc. Am. Math Soc. **3,** 643–646 (1952).

[9] H. F. Weinberger, An optimum problem in the Weinstein method for eigenvalues, Pacific J. Math. **2,** 413–418 (1952).

[10] A. Weinstein, Separation theorems for the eigenvalues of partial differential equations, Reissner Anniversary Volume, p. 405–416 (1949).

[11] A. Weinstein, Quantitative methods in Sturm-Liouville theory, Proc. of Symposium on Spectral Theory and Differential Problems, p. 345–352 (Oklahoma A. and M. College, 1951).

[12] H. F. Weinberger, The connection between one-dimensional eigenvalue problems with interior boundary conditions, Bul. Am. Math. Soc. **57,** 182 (1951).

[13] C. Arf, On the methods of Rayleigh-Ritz-Weinstein. Proc. Am. Math. Soc. **3,** 223–232 (1952).

12. Determination of Eigenvalues and Eigenvectors of Matrices

Magnus R. Hestenes[1]

12.1. Introduction

The problem at hand is that of finding a number λ and an n-dimensional vector $y \neq 0$ such that $Ay=\lambda y$, where A is an n-dimensional square matrix. The vector y is called an *eigenvector* of A and the number λ is called an *eigenvalue* of A. Unless otherwise expressly stated, we shall be concerned with a symmetric matrix A over the field of reals. The extension of these results to Hermitian matrices over the complex field is immediate.

In the present paper we describe certain methods that have been studied at the Institute for Numerical Analysis, National Bureau of Standards. Our experiments indicate that these methods can be used effectively in computations for symmetric or Hermitian matrices. The Institute is sponsoring a program of study of methods of finding eigenvalues and eigenvectors of an arbitrary matrix.

12.2. The Power Method

One of the best-known methods of finding eigenvalues is the power method. In this method a sequence of vectors $\{x_i\}$ is constructed by the formula

$$x_{i+1}=\alpha_i A x_i \qquad (i=0,1,2,\ldots).$$

The initial vector x_0 is arbitrary, and the numbers α_i are scale factors. In this method one does not need to restrict A to be real and symmetric. If there is a unique eigenvalue λ of maximum absolute value the sequence $\{x_i\}$ will converge to an eigenvector y corresponding to λ. The ratios of corresponding nonzero components of x and Ax will converge to λ.

We shall not dwell further on the power method. Its properties are well known. The method can be modified in many ways so as to speed up convergence. It is closely related to the gradient method, which will be discussed in the next section.

12.3. The Gradient Method [1]

The gradient method is based on the fact that if the equation $Ax=\lambda x$, $(x\neq 0)$, holds, then λ is given by the formula $\lambda=\mu(x)$, where

$$\mu(x)=\frac{x^*Ax}{|x|^2}. \tag{1}$$

Here x^* denotes the transpose of x. If A is real and symmetric, as we shall assume, the critical points of $\mu(x)$ are eigenvectors and the critical values are eigenvalues. In particular the maximum of $\mu(x)$ is the greatest eigenvalue and the minimum of $\mu(x)$ is the least eigenvalue. This suggests that these extreme eigenvalues can be obtained by finding the maximum and minimum of $\mu(x)$. The direction of steepest ascent is given by the vector

$$\xi=Ax-\mu(x)x, \tag{2}$$

which we shall call the *gradient* of $\mu(x)$ *at* x, or more simply the *gradient at* x. It is clear that a vector $x\neq 0$ is an eigenvector if, and only if, $\xi=0$. Thus, the magnitude of ξ relative to the magnitude of x can be taken as a measure of the deviation of x from an eigenvector.

In order to obtain the maximum of $\mu(x)$ by use of the gradient method, we use the iteration

$$x_{i+1}=x_i+\alpha_i\xi_i \qquad (i=0,1,2,\ldots). \tag{3}$$

[1] National Bureau of Standards, Los Angeles, Calif., and University of California, Los Angeles.

where ξ_i is the gradient of $\mu(x)$ at x_i. It can be shown that the scalars α_i can be chosen so that the inequality

$$\mu(x_{i+1})-\mu(x_i)\geqq c|\xi_i|^2 \tag{4}$$

holds, where c is a positive number independent of i. In particular, this is the case when α_i is on the range

$$0<\delta\leq\alpha_i\leq\frac{2}{M}-\delta, \tag{5}$$

where M is the spread of the eigenvalues. If α_i is chosen so that (5) holds, then the sequence $\{\mu(x_i)\}$ will converge to the greatest eigenvalue $\lambda_{\max}$ unless the initial vector x_0 is orthogonal to the eigenvectors corresponding to $\lambda_{\max}$. Moreover, the sequence $\{x_i\}$ will converge to the corresponding eigenvector. If α_i is replaced by $-\alpha_i$ in (3), the minimum is obtained in place of the maximum. In this event the inequality (4) takes the form

$$\mu(x_i)-\mu(x_{i+1})\geq c|\xi_i|^2.$$

There are three methods of choosing the parameter α_i which have proved to be practical.

The first method consists of assigning a fixed value to α_i. Normally we seek to select α_i to be on the range (5). Inasmuch as the spread $M=\alpha_{\max}-\alpha_{\min}$ is not known beforehand, one might expect that it would be difficult to make a suitable choice of α_i. However, this is not the case. If the largest eigenvalue is sought the computer adjusts α_i so that $\mu(x_{i+1})>\mu(x_i)$. If at any step this inequality fails to hold, the computer knows that α_i is too large by at least the factor two [1]. Thus, a suitable choice of α can be determined by watching the growth of $\mu(x)$. An experienced computer will vary α from time to time in order to accelerate the convergence. We have found that interspersing a few large values of α with smaller ones is very effective.

The second method is to select α_i at the ith stage so that $\mu(x_i+\alpha\xi_i)$ is a maximum. In this event α_i is given by the formula

$$\alpha_i=\frac{2}{-s_i+\sqrt{s_i^2+4t_i^2}},$$

where

$$s_i=\mu(\xi_i)-\mu(x_i),\quad t_i=\frac{|\xi_i|}{|x_i|}. \tag{6}$$

When t_i is small, as it will be when x_i is a good estimate of the solution, then $\alpha_i=1/|s_i|$ is a good estimate of the optimum value of α_i. In any event, it is an over estimate of this optimum value. It has been our experience that this method converges too slowly. Normally accelerations are needed. We have found that this method is not as good as the one described above.

The third method is to choose α_i to be a fraction of the optimum value described in the preceding paragraph. For example, one can choose $\alpha_i=\beta/|s_i|$, where s_i is given by (6) and β is a fixed number on the range $0<\beta<1$. In particular, the values $\beta=.7$, $\beta=.8$, $\beta=.9$ have been found to be effective in selected examples. This method has the property of being self-accelerating in the sense that at relatively frequent, but irregular, intervals large corrections are made. The experiments that we have carried out to date seem to indicate that this method is superior to those discussed above, including the power method.

The connection between the gradient method and the power method can be seen if we write the iteration (3) in the form

$$x_{i+1}=\alpha_i(A-c_iI)x_i,$$

where

$$c_i=\mu(x_i)-\frac{1}{\alpha_i}.$$

In the power method, c_i is held fast. In the gradient method, c_i is determined at each stage. It is clear that a wise choice of c_i will greatly accelerate the process. It is for this reason that the gradient method appears to be preferable to the power method.

12.4. A Generalization of the Gradient Method [2]

The gradient method described in the previous section can be looked upon from a slightly different point of view. Let $\{x,Ax,. . .,A^{k-1}x\}$ be the class of vectors that are linear combinations of $x, Ax, \ldots, A^{k-1}x$. In the gradient method the vector x_{i+1} is chosen in the subspace $\{x_i,Ax_i\}$, so that the inequality

$$\mu(x_{i+1})-\mu(x_i)\geq c|\xi_i|^2 \tag{7}$$

holds, where c is a small positive number independent of i. An obvious extension of this procedure is to select x_{i+1} in the class $\{x_i,Ax_i,. . .,A^{k-1}x_i\}$, where k is fixed. One keeps the condition (7). If one seeks the minimum eigenvalue in place of the maximum, one uses the inequality

$$\mu(x_i)-\mu(x_{i+1})\geq c|\xi_i|^2$$

in place of (7).

The difficulty encountered in this procedure is that of selecting a suitable vector x_{i+1}. For example, one might select x_{i+1} so as to maximize (or minimize) the function $\mu(x)$. This would involve solving the characteristic equation $P_k(\lambda)=0$ of A on the space $\{x_i,Ax_i, . . .,A^{k-1}x_i\}$. Methods for doing this will be described in the sections that follow.

A theoretical discussion of the method just described can be found in a recent paper by Karush [2].

12.5. Iterative Methods of Generating the Characteristic Polynomial [3, 4, 5].

The present section will be devoted to describing an iterative method of generating the characteristic polynomial of A. For convenience, we shall assume that A has n distinct eigenvalues. This is not an essential restriction and is made to simplify our discussions. The method we shall describe is a generalization of one due to Lanczos [3].

Let B be a positive symmetric matrix that commutes with the given symmetric matrix A. We suppose that a vector x_0 has been chosen so that the vectors $x_0, Ax_0, . . ., A^{n-1}x_0$ are linearly independent. We select x_1 so that

$$x_1=a_0Ax_0-b_0x_0, \tag{8}$$

where a_0 is a scale factor, and b_0 is chosen so that $x_1^*Bx_0=0$. The formula for b_0 is

$$b_0=a_0\frac{x_0^*BAx_0}{x_0^*Bx_0}.$$

Having chosen the vectors $x_0, . . ., x_i$ $(i\neq 1)$, we select x_{i+1} by the formula

$$x_{i+1}=a_iAx_i-b_ix_i-c_{i-1}x_{i-1}, \tag{9}$$

where a_i is a nonzero scale factor and

$$b_i=a_i\frac{x_i^*BAx_i}{x_i^*Bx_i}$$

$$c_{i-1}=a_i\frac{x_{i-1}^*BAx_i}{x_{i-1}^*Bx_{i-1}}=\frac{a_i}{a_{i-1}}\frac{x_i^*Bx_i}{x_{i-1}^*Bx_{i-1}}. \tag{10}$$

The vectors $x_0, . . ., x_{n-1}$ obtained in this manner satisfy the relation

$$x_i^*Bx_j=0, \qquad i\neq j. \tag{11}$$

This can be proved by induction (see reference [4]). We have already seen that $x_0^*Bx_1=0$. Suppose this relation holds when $i<j\leq k$. It is easy to see from the definition of b_k and c_{k-1} that the relation holds with $i=k$, $k-1$, and $j=k+1$. Suppose now that $0<i<k-1$. Then

$$\begin{aligned} x_i^*Bx_{k+1}&=a_kx_i^*BAx_k-b_ix_i^*Bx_k-c_{i-1}x_i^*Bx_{k-1}\\ &=a_kx_k^*BAx_i\\ &=\frac{a_k}{a_i}x_k^*B(x_{i+1}+bx_i+c_{i-1}x_{i-1})=0. \end{aligned}$$

A similar argument holds for $i=0$.

From the definition of x_i it is clear that x_i is of the form $x_i=P_i(A)x_0$, where $P_i(\lambda)$ is a polynomial of degree i in λ. It is clear from (11) that $x_n=P_n(A)x_0=0$. Hence

$$P_n(A)x_i=P_n(A)P_i(A)x_0=P_i(A)P_n(A)x_0=0.$$

It follows that $P_n(A)=0$. The polynomial $P_n(\lambda)$ is accordingly the characteristic polynomial of A. From the relations (8) and (9) it follows that the polynomials $P_i(\lambda)$ satisfy the relations

$$P_0(\lambda)=1, \qquad P_1(\lambda)=a_0\lambda-b_0, \qquad P_{i+1}(\lambda)=(a_i\lambda-b_i)P_i(\lambda)-c_{i-1}P_{i-1}(\lambda). \tag{12}$$

Similarly, the derivatives of $P_i(\lambda)$ satisfy the relations

$$P_0'(\lambda)=0, \qquad P_1'(\lambda)=a_0, \qquad P_{i+1}'(\lambda)=(a_i\lambda-b_i)P_i'(\lambda)-c_{i-1}P_{i-1}'(\lambda)+a_iP_i(\lambda). \tag{13}$$

By using (12) and (13) one can obtain the values of $P_n(\lambda)$ and $P_n'(\lambda)$ for a given number λ without obtaining an explicit formula for P_n. We can accordingly apply Newton's formula

$$\lambda_1=\lambda-\frac{P_n(\lambda)}{P_n'(\lambda)}$$

to compute the zeros of P_n, that is, the eigenvalues of A. Having found an eigenvalue λ, a corresponding eigenvector y can be obtained by the formula

$$y=h_0x_0+\ . \ . \ . \ +h_{n-1}x_{n-1},$$

where

$$h_i=\frac{P_i(\lambda)}{x_i^*Bx_i} \qquad (i=0,1,\ .\ .\ .,\ n-1).$$

The case $B=I$ is the one discussed by Lanczos [3]. From a theoretical point of view nothing is gained by using a more general matrix B. This follows from the fact that the vector $\bar{x}_0=B^{\frac{1}{2}}x_0$ will generate the same polynomials (using $B=I$) as those given above. Whether or not there is a computational advantage remains to be seen.

The iteration described above normally can be put in a somewhat different form. We consider only the case $B=I$. The new form is given by the system

$$\begin{aligned} x_0&=y_0 \\ x_{i+1}&=\alpha_ix_i+\beta_iAy_i \\ y_{i+1}&=\gamma_iy_i+\delta_ix_{i+1}. \end{aligned} \tag{14}$$

Here β_i and δ_i are nonzero scale factors and

$$\alpha_i=-\beta_i\frac{x_i^*Ay_i}{|x_i|^2}, \qquad \gamma_i=-\delta_i\frac{y_i^*Ax_{i+1}}{y_i^*Ay_i}.$$

In the unusual case when $y_i^*Ay_i=0$ for some $i<n-1$ this system fails. It is not difficult to show that the relations

$$x_i^*x_j=0, \quad y_i^*Ay_j=0, \qquad (i\neq j) \tag{15}$$

hold. Moreover, upon eliminating the y's in (14) one obtains the formulas

$$x_i=\beta_0Ax_0+\alpha_0x_0$$

$$x_{i+1}=a_iAx_i-b_ix_i-c_{i-1}x_{i-1}, \qquad (i>0).$$

vhere

$$a_i=\beta_i\gamma_i, \qquad b_i=-\alpha_i-\delta_i\beta_i/\beta_{i-1}, \qquad c_{i-1}=\alpha_{i-1}\delta_i\beta_i/\beta_{i-1}.$$

n a similar manner it can be seen that the relations

$$y_1=\beta_0\gamma_0Ay_0+(\delta_0+\alpha_0\gamma_0)y_0$$

$$y_{i+1}=a'_iAy_i-b'_iy_i-c'_{i-1}y_{i-1}$$

ıold where

$$a'_i=\gamma_i\beta_i, \qquad b'_i=-\delta_i-\alpha_i\gamma_i/\gamma_{i-1}, \qquad c'_{i-1}=\delta_{i-1}\alpha_i\gamma_i/\gamma_{i-1}.$$

Chus, it is seen that the x's and the y's have the properties described at the beginning of this section. The haracteristic polynomial $P_n(\lambda)$ can be obtained from the formulas

$$P_0=Q_0=1$$

$$P_{i+1}=\alpha_iP_i+\beta_i\lambda Q_i$$

$$Q_{i+1}=\delta_iP_{i+1}+\delta_iQ_i$$

nd its derivative P'_n by the formulas

$$P'_0=Q'_0=0$$

$$P'_{i+1}=\alpha_iP'_i+\beta_i\lambda Q'_i+\beta_iQ_i$$

$$Q'_{i+1}=\gamma_iP'_{i+1}+\delta_iQ'_i.$$

Consequently, Newton's methods can be applied as before to obtain the zeros of $P_n(\lambda)$.

The methods described in this section can be readily generalized to the case in which A is not real nd symmetric.

12.6. A Block Method

We shall describe briefly a method that has been used successfully on a matrix which was too large o be handled by the gradient method with the equipment on hand. We shall call this the block method. t can be described as follows:

(1) Partition x into two parts, $x=(y,z)$, where y denotes k components of x and z the remaining $-k$ components of x. The function $\mu(x)$ described in section 12.3 is a function $\mu(y,z)$ of y and z.

(2) Holding $z=z_1$ fast, select y_1 and a scalar α_1 such that $\mu(y,\alpha z_1)$ has a maximum value. If $\alpha\neq 0$, e select y_1 and α_1 so that $\alpha_1=1$. A vector $x_1=(y_1, \alpha_1z_1)$ is obtained.

(3) Iterate steps (1) and (2), using a new partition for each iteration. In each case the components o be held fast are assigned the values of the corresponding components of the vector x_{i-1} determined ı step (2) of the preceding iteration.

When this procedure is carried out in a cyclic manner, a sequence of vectors $\{x_i\}$ is obtained which onverge, if suitably normalized, to an eigenvector corresponding to the highest eigenvalue.

The procedure just described is analogous to that given in section 12.4. In each case $\mu(x)$ is maxiıized (or minimized) successively on linear subspaces $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, . . . of our space. In this section, he subclass $\boldsymbol{B}_i$ is determined by x_i and the partition of x, whereas in section 12.4 the space $\boldsymbol{B}_i$ was deermined by the vectors x_i, Ax_i, . . ., $A^{k-1}x_i$.

The method just described appears to be well adapted for machine computation.

12.7. *The Case $Ax=\lambda Bx$* [6]

We now turn to the problem of finding a solution of the system

$$Ax=\lambda Bx, \tag{16}$$

where A and B are n-dimensional square matrices. We consider the problem in the domain of complex numbers, and assume that B is nonsingular. The matrices A and B need not be Hermitian. If λ is chosen so that (16) has a nonnull solution x, then λ is an *eigenvalue* and x an *eigenvector.*

Let C, D, $K(\sigma)$ be the matrices

$$C=B^*HA, \qquad D=B^*HB, \qquad K(\sigma)=(A-\sigma B)^*H(A-\sigma B), \tag{17}$$

where H is a positive Hermitian matrix. It is clear that D is a positive Hermitian matrix and that $K(\sigma)$ is a nonnegative Hermitian matrix. In fact, $K(\sigma)$ is positive whenever $A-\sigma B$ is nonsingular, that is, if σ is not an eigenvalue of the system (16).

The results to be described in this section are centered around the two functions

$$f(x,\sigma)=\frac{x^*K(\sigma)x}{x^*Dx}, \qquad \mu(x)=\frac{x^*Cx}{x^*Dx}. \tag{18}$$

These functions are connected by the relation

$$f(x,\sigma)=f(x,\mu(x))+|\sigma-\mu(x)|^2. \tag{19}$$

From this result it is seen that $f(x,\sigma)=0$ if, and only if, $\sigma=\mu(x)$ and $f(x,\mu(x))=0$. Consequently, a vector $x\neq 0$ and a number λ is a solution of (19) if, and only if, $f(x,\lambda)=0$. Moreover, $\lambda=\mu(x)$. Thus, a vector x is an eigenvector if, and only if, $f(x,\mu(x))=0$. For an eigenvector x the formula (19) takes the form

$$f(x,\sigma)=|\sigma-\mu(x)|^2. \tag{20}$$

Consider now a fixed number σ and let r_1 and r_2 be the greatest and the least nonnegative numbers such that $r_1^2\leq f(x,\sigma)\leq r_2^2$. For an eigenvector x we have by (20) the relation

$$r_1\leq|\sigma-\mu(x)|\leq r_2.$$

It follows that there is no eigenvalue within the circle Γ_1 of radius r_1 about σ and no eigenvalue exterior to the circle Γ_2 of radius r_2 about σ. By constructing the circles Γ_1 and Γ_2 for various values of σ one determines domains in which eigenvalues lie and in which they do not lie. If $C^*=C$, it can be shown that there is an eigenvalue on each of the circles Γ_1 and Γ_2. More generally, if the system (16) has n-linearly independent eigenvectors, it is always possible to choose H so that there is an eigenvalue on each of these circles.

The remarks made above suggest the following procedure for finding eigenvalues and eigenvectors. Select a number σ_1 and determine x_1 such that $f(x,\sigma_1)$ has a minimum value at x_1. Select $\sigma_2=\mu(x_1)$, and determine x_2 such that $f(x,\sigma_2)$ has a minimum value at x_2. Having chosen x_{i-1}, set $\sigma_i=\mu(x_{i-1})$, and select x_i so that $f(x,\sigma_i)$ attains its minimum at x_i. This procedure will determine a sequence x_i such that $\mu(x_i)$ will converge to an eigenvalue whenever $f(x_i,\sigma_i)$ tends to zero. This method has been carried out successfully in a special case. Whether or not it is a good method for computation is yet to be determined.

References

[1] M. R. Hestenes and W. Karush, Method of gradients for the calculation of the characteristic roots and vectors of a real symmetric matrix, J. Research NBS **47**, 45 (1951) RP2227.

[2] W. Karush, An iterative method for finding characteristic vectors of a symmetric matrix, Pacific J. Math. **I**, 23 (June 1951).

[3] Cornelius Lanczos, Iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Research NBS **45**, 255 (1950) RP2133.

[4] M. R. Hestenes, Iterative methods for solving linear equations, publication pending.

[5] J. B. Rosser, C. Lanczos, M. R. Hestenes, and W. Karush, The separation of close eigenvalues of a real symmetric matrix, J. Research NBS **47**, 291 (1951) RP 2256.

[6] M. R. Hestenes and W. Karush, The solutions of $Ax=\lambda Bx$, J. Research NBS **47**, 471 (Dec. 1951), RP 2275.

13. New Results in the Perturbation Theory of Eigenvalue Problems

F. Rellich [1]

13.1. Introduction

Let us consider a self-adjoint eigenvalue problem $A\varphi=\lambda\varphi$, where $A=A(\epsilon)$ depends analytically on , perturbation parameter: $A(\epsilon)=A_0+\epsilon A_1+\ .\ .\ .$, the simplest case being $A(\epsilon)=A_0+\epsilon A_1$. One expects hat eigenvalues and eigenfunctions of the disturbed operator $A(\epsilon)$ will depend analytically on ϵ for mall $|\epsilon|$ thus

$$\left.\begin{aligned}\lambda(\epsilon)&=\lambda^{(0)}+\epsilon\lambda^{(1)}+\ .\ .\ .\\ \varphi(\epsilon)&=\varphi^{(0)}+\epsilon\varphi^{(1)}+\ .\ .\ .\end{aligned}\right\}. \tag{1}$$

'he computation of $\lambda^{(1)}$, $\varphi^{(1)}$, . . . yields the successive approximations of the unknown $\lambda(\epsilon)$, $\varphi(\epsilon)$.

Example 1. One of the first problems treated in this way in the literature is Lord Rayleigh's roblem of the vibrating string of small stiffness. In a slightly generalized form this is

$$-u''+q(x)u+\epsilon u''''=\lambda u, \quad 0\leqq x\leqq 1, \quad u(0)=u''(0)=u(1)=u''(1)=0.$$

think nobody will expect that compared with the undisturbed term $-u''+q(x)u$ the perturbation u'''' is small enough to allow a series of the kind (1) with $\epsilon\neq 0$.

Example 2. In quantum mechanics the perturbation method was first used by E. Schrödinger, 'ho treated the eigenvalue equation of the Stark effect, i. e.,

$$-(u_{xx}+u_{yy}+u_{zz})-\frac{2Z}{r}u+\epsilon xu=\lambda u, \quad -\infty<x,y,z<+\infty, \quad r=(x^2+y^2+z^2)^{1/2}$$

n this example, it is perhaps less obvious whether or not the perturbation term ϵxu is small enough for small $|\epsilon|$) as to guarantee the development (1). But from the behavior of the potential energy -$2Z/r+\epsilon x$ one may conjecture that for $\epsilon\neq 0$ the problem has a continuous spectrum without point eigen-alues and that means that (1) is not true, at least not in the usual sense.

Example 3. The wave equation of the nonharmonic oscillator is given by

$$-u''+(x^2+\epsilon x^4)u=\lambda u, \quad -\infty<x<\infty$$

vith an additional boundary condition at $x=+\infty$ and $x=-\infty$ in the case $\epsilon<0$ in order to have a self-djoint problem). This problem has a pure point-spectrum for every (real) value of ϵ. But for $\epsilon\geqq 0$ nd for $\epsilon<0$ the spectra of the problem are quite different. In the first case we have $\lim_{n\to\infty}\lambda_n(\epsilon)=+\infty$; 1 the second case the spectrum consists of two sequences $\lambda_n(\epsilon)$, $\mu_n(\epsilon)$ of eigenvalues with $\lim_{n\to\infty}\lambda_n(\epsilon)=+\infty$, $\lim_{\to\infty}\mu_n(\epsilon)=-\infty$.

This difference is well known for the mechanical analogue, i. e., Duffing's vibration problem with the onlinear restoring force $-2x-4\epsilon x^3$). Again it is doubtful whether the perturbation ϵx^4u is small nough as to expect convergent power series (1).

In each of these examples it is easy to compute $\lambda^{(1)}$ by the well-known perturbation procedure. ut it is not easy to say what is the sense of $\lambda^{(0)}+\epsilon\lambda^{(1)}$ as an approximation, if there is a sense in it at l.

[1] University of Göttingen, Germany.

The question, what is a "small" perturbation, is answered in a satisfactory way for several types of eigenvalue problems. The answer is not a unique one. A perturbation can be small with respect to the isolated point eigenvalues of the undisturbed operator but not small with respect to its continuous spectrum. Certainly there are a lot of reasonable questions for which the perturbation theory as developed until now gives no answer, but, in spite of that, I think the perturbation theory of eigenvalue problems is in a better situation than its elder brother, the perturbation theory of classical mechanics, which, e. g., until now has no other characterization for small perturbations than smallness of the perturbation parameter.

I gave a survey of the known mathematical results in the perturbation theory of eigenvalue problems at the International Congress of Mathematicians in Cambridge (Mass.) last year.[2] Since that time two new results were obtained, which I will describe in what follows. Both of them are due to E. Heinz.[3]

13.2. The Dependence on the Perturbation Parameter for the Resolution of the Identity

To describe the spectrum of a self-adjoint operator Hilbert introduced the resolution of identity E_λ, $-\infty<\lambda<\infty$. Instead of a definition, I give an example: Let

$$Au=-u'',\quad -l<x<l,\quad u(-l)=u(l),\quad u'(-l)=u'(l).$$

Then the resolution of the identity of this operator is given by

$$E_\lambda u=\sum_{\lambda_n<\lambda}(\varphi_n,u)\varphi_n,\quad \lambda_n=n^2\pi^2/l^2,\quad \varphi_n=(2l)^{-1/2}e^{in\pi x/l},\quad (\varphi_n,u)=\int_{-l}^{+l}\overline{\varphi_n(x)}u(x)dx,\quad n=0,\pm1\pm2,\ldots.$$

If A depends on ϵ we shall have $E_\lambda=E_\lambda(\epsilon)$. Now we make the following assumptions about $A=A(\epsilon)$: *In a subspace* ***A*** *of a Hilbert space* ***H*** *the operator A_0 is self-adjoint and A_1, A_2, . . . are Hermitian operators in* ***A***. *For each u of A we require $A(\epsilon)u=A_0u+\epsilon A_1u+\ldots$, where the power series at the righthand side converges in the sense given by the norm $||u||=(u,u)^{1/2}$ defined in the Hilbert space. Then it can be proved that two constants a and k exist such that*

$$||A_\nu u||\leqq ak^\nu(||A_0u||+||u||),\qquad \nu=0,1,2,\ldots \tag{2}$$

for all u of ***A*** *and that for small ϵ the operator $A(\epsilon)$ is self-adjoint in* ***A***.

For differential operators, however, it cannot always easily be decided whether or not an operator A_1 (or A_2, A_3, . . .) can be defined as an Hermitian operator in the subspace ***A*** in which A_0 is self-adjoint It is therefore more convenient to suppose that A_0 is essentially self-adjoint[4] in a subspace ***D*** of ***A*** which, in general, will be smaller than ***A*** and will allow to see immediately that A_1, A_2, . . . can be defined as Hermitian operators. Now the existence of two constants a, k for which the inequalities (2) hold can no longer be proved but are an additional assumption. With the aid of this assumption one can prove that $A(\epsilon)=A_0+\epsilon A_1+\ldots$ defined in ***D*** can be extended by closure to an operator self-adjoint in $\boldsymbol{A}\supseteq\boldsymbol{D}$. Thus the first definition of $A(\epsilon)$ in ***A*** is regained and we shall use it further in this section. (The assumptions of this definition are not fulfilled in our three examples of the introduction (with $A_2=A_3=\ldots=0$) but they are fulfilled for the second and the third example if the independent variables are restricted to a "finite box").

The question arises whether such an operator $A(\epsilon)=A_0+\epsilon A_1+\ldots$, self-adjoint in ***A*** for small $|\epsilon|$ has an $E_\lambda(\epsilon)$ that is regular with respect to ϵ, $E_\lambda(\epsilon)=E_\lambda^{(0)}+\epsilon E_\lambda^{(1)}+\ldots$ (The operators $E_\lambda(\epsilon)$, $E_\lambda^{(0)}$, E_λ^1, . . . being bounded operators, the meaning of this series is obvious).

B. v. Nagy [5] proved the following important theorem. *If the spectrum of the undisturbed operator A_0 is empty in the two intervals $\lambda_0-d<\lambda<\lambda_0+d$ and $\mu_0-d<\lambda<\mu_0+d$ then $E_{\lambda_0}(\epsilon)-E_{\mu_0}(\epsilon)=P_0+\epsilon P_1+\ldots$,*

[2] F. Rellich, Störungstheorie der Spektralzerlegung, Proc. Intern. Congr. Math. Cambridge (Mass.) (1950).

[3] E. Heinz, Beiträge zur Störungstheorie der Spektralzerlegung, Math. Ann. **123**, 415–438 (1951).

[4] A_0 is essentially self-adjoint in ***D*** if it is self-adjoint in $\boldsymbol{A}\supseteq\boldsymbol{D}$ and if to each u of ***A*** a sequence u_n of ***D*** can be found with $||u-u_n||\to0$ and $||A_0(u-u_n)||\to0$, $n\to\infty$, A_0 in ***A*** is called the closure of A_0 in ***D***.

[5] B. de Sz. Nagy, Perturbations des transformations autoadjointes des l'espace de Hilbert, Comment. Math. Helv. **19**, 347–366 (1946).

vhere $P_0, P_1, \ldots$ are bounded operators and the series is convergent for small values of $|\epsilon|$. E. Heinz liscovered that even $E_{\lambda_0}(\epsilon)=Q_0+\epsilon Q_1+\ldots$ (with bounded $Q_0, Q_1, \ldots$ and converging series) is true rovided that the spectrum of A_0 is empty in $\lambda_0-d<\lambda<\lambda_0+d$. This theorem of Heinz contains the heorem of v. Nagy, but it is much further reaching. I shall not give the proof here, but in section 13.4 shall prove an inequality that is used by Heinz as an essential tool in his proof.

13.3. Convergent Sequences of Operators

Several years ago I proved the following theorem. *Let A in $\boldsymbol{A}$ and $A^{(n)}$ in $\boldsymbol{A}^{(n)}$, $n=1, 2, \ldots$ be elf-adjoint operators, $\boldsymbol{A}$, $\boldsymbol{A}^{(n)}$ subspaces of a Hilbert space $\boldsymbol{H}$. Let the subspace $\boldsymbol{D}$ be contained in each f the spaces $\boldsymbol{A}, \boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)}, \ldots$ and let A in $\boldsymbol{D}$ be essentially self-adjoint. Finally,* $\lim_{n\to\infty} ||(A^{(n)}-A)u||=0$ *or each u of $\boldsymbol{D}$. If λ_0 is not a point-eigenvalue of A, then* $\lim_{n\to\infty} ||(E_{\lambda_0}^{(n)}-E_{\lambda_0})f||=0$ *for all f of $\boldsymbol{H}$, where $E_\lambda^{(n)}$ resp. E_λ are the resolutions of identity of $A^{(n)}$ resp. A.* This theorem is rather general. It even overs Rayleigh's example mentioned above. Indeed, call $\boldsymbol{D}$ the space of all functions $u(x)$ with coninuous fourth derivatives in $0\leqq x\leqq 1$ and $u(0)=u''(0)=u(1)=u''(1)=0$. The operator $-(d^2/dx^2)+q(x)$ s essentially self-adjoint in $\boldsymbol{D}$, we call A its closure. The operator $-(d^2/dx^2)+q(x)+\epsilon_n(d^4/dx^4)$ with $\epsilon_n\neq 0$, $\lim_{n\to\infty} \epsilon_n=0$, is again essentially self-adjoint in $\boldsymbol{D}$, we call its closure $A^{(n)}$. Obviously

$$||(A^{(n)}-A)u||=|\epsilon_n|\left(\int_0^1 \left|\frac{d^4u}{dx^4}\right|^2 dx\right)^{\frac{1}{2}} \to 0,\ n\to\infty.$$

hus $\lim_{n\to\infty} ||(E_\lambda^{(n)}-E_{\lambda_0})f||=0$ if λ_0 is no eigenvalue of A. Because of the fact that this limit-relation n general is not true uniformly for all f of $\boldsymbol{H}$ with $||f||\leqq 1$ the statement established by this relation s a very weak one.

In order to obtain uniform convergence we replace the condition

$$\lim_{n\to\infty} ||(A^{(n)}-A)u||=0,\ u\,\epsilon\,\boldsymbol{D}$$

y the stronger condition

$$||(A^{(n)}-A)u||\leqq \eta_n\{||u||+||Au||\}$$

or all u of $\boldsymbol{D}$, where η_n is a sequence of numbers with $\lim_{n\to\infty} \eta_n=0$. (In Rayleigh's example this stronger ondition is obviously not satisfied).

The question is whether this stronger condition guarantees $\lim_{n\to\infty} ||(E_{\lambda_0}^{(n)}-E_{\lambda_0})f||=0$, uniformly for all of $\boldsymbol{H}$ with $||f||\leqq 1$ in the case that the spectrum of A is empty in an interval $\lambda_0-d<\lambda<\lambda_0+d$. Heinz ave the first proof for this statement. Again an important tool of this proof is the inequality of the next ection.

13.4. Heinz's inequality

Let $\boldsymbol{D}$ be a subspace of a Hilbert space, A in $\boldsymbol{D}$ an essentially self-adjoint operator and Q in $\boldsymbol{D}$ an Iermitian operator. Further, $0\leqq(u,Au)$ and $(Qu, Qu)\leqq(Au, Au)$ for all u of $\boldsymbol{D}$. Then

$$|(u,Qu)|\leqq(u,Au) \qquad \textit{for all } u \textit{ of } \boldsymbol{D}. \tag{3}$$

his inequality is the simplest among much more general inequalities obtained by Heinz,[6] but it is suffiient for the purposes mentioned in section 13.2 and 13.3. In what follows we shall give a simple proof f (3). From (3) follows $|(u,Qv)|^2\leqq(u,Au)(v,Av)$ for all u,v of $\boldsymbol{D}$ in the well known way.

In the special case of an n-dimensional Hilbert space and a nonnegative Q, i. e., $(u,Qu)\geqq 0$ for u n $\boldsymbol{D}$, the inequality is known. In fact, if we put $R=Q^2$, $S=A^2$, we have $(u,Ru)\leqq(u,Su)$, and thus the nequality $(u,R^{1/2}u)\leqq(u,S^{1/2}u)$ is well known from Loewner's theory of monotone functions of matrices.

[6] See footnote 3.

A simple proof [7] (for the n-dimensional space) runs as follows. Let λ be an eigenvalue of the Hermitian operator $A-Q$, thus $(A-Q)\varphi=\lambda\varphi, \|\varphi\|=1$. Hence

$$\lambda((A+Q)\varphi,\varphi)=((A+Q)\varphi,(A-Q)\varphi)=(A\varphi,A\varphi)-(Q\varphi,Q\varphi)+(Q\varphi,A\varphi)-(A\varphi,Q\varphi).$$

Taking the real part of this equation, we get

$$\lambda((A+Q)\varphi,\varphi)=(A\varphi,A\varphi)-(Q\varphi,Q\varphi)\geqq 0.$$

Adding $\lambda((A-Q)\varphi,\varphi)\geqq\lambda^2$, we have $2\lambda(A\varphi,\varphi)\geqq\lambda^2$, and therefore $\lambda\geqq 0$ because of $(A\varphi,\varphi)\geqq 0$. However if every eigenvalue of $A-Q$ is positive, or zero, then $(u,Au)\geqq(u,Qu)$. Replacing Q by $-Q$, we get $(u,Au)\geqq-(u,Qu)$; hence $(u,Au)\geqq|(u,Qu)|$.

With some carefulness the idea of this proof can be used in the Hilbert space of not finite dimension.

The closure of A in $\boldsymbol{D}$ is a self-adjoint operator A in $\boldsymbol{A}\supset\boldsymbol{D}$. Because of $(Qu,Qu)\leqq(Au,Au)$ the operator Q can be extended to an Hermitian operator in $\boldsymbol{A}$. The operator $A-Q$ is in general not self-adjoint in $\boldsymbol{A}$. Therefore, we introduce $A-kQ$, where k is a fixed number, $-1<k<1$. This operator is self-adjoint in $\boldsymbol{A}$. Indeed $A-kQ-i=(1-kR)(A-i)$ with $R=Q(A-i)^{-1}$. The operator R is defined in $\boldsymbol{H}$ and we have $\|Rf\|=\|Q(A-i)^{-1}f\|$ or, with $u=(A-i)^{-1}f,(A-i)u=f$, $\|Qu\|\leqq\|Au\|\leqq\|f\|$, hence $\|Rf\|\leqq\|f\|$. Therefore, $(A-kQ-i)^{-1}=(A-i)^{-1}(1+kR+k^2R^2+\ldots)$ is a bounded operator. The same being true for $(A-kQ+i)^{-1}$ we have proved, that $A-kQ$ is self-adjoint in $\boldsymbol{A}$.

Let λ be a point of the spectrum of $A-kQ$. Then the spectral space of this operator belonging to the spectral interval $\lambda-1/n$ to $\lambda+1/n$, $n=1, 2, \ldots$, is not empty; it contains at least one element φ_n, thus

$$\left(E_{\lambda+\frac{1}{n}}-E_{\lambda-\frac{1}{n}}\right)\varphi_n=\varphi_n, \qquad \|\varphi_n\|=1.$$

It follows

$$(A-kQ)\varphi_n-\lambda\varphi_n=\omega_n \quad \text{if} \quad \omega_n=\int_{\lambda-\frac{1}{n}}^{\lambda+\frac{1}{n}}(\mu-\lambda)dE_\mu\varphi_n.$$

We find

$$\begin{aligned}\lambda((A+kQ)\varphi_n,\varphi_n)&=((A+kQ)\varphi_n,(A-kQ)\varphi_n)-((A+kQ)\varphi_n,\omega_n)\\&=(A\varphi_n,A\varphi_n)-k^2(Q\varphi_n,Q\varphi_n)-\mathrm{Re}\{((A+kQ)\varphi_n,\omega_n)\}\\&\geqq-|((A+kQ)\varphi_n,\omega_n)|\\&\geqq-\|(A+kQ)\varphi_n\|\cdot\|\omega_n\|.\end{aligned}$$

Adding

$$\lambda((A-kQ)\varphi_n,\varphi_n)=\lambda^2+\lambda(\omega_n,\varphi_n)\geqq\lambda^2-|\lambda|\ \|\omega_n\|,$$

we have

$$2\lambda(A\varphi_n,\varphi_n)\geqq\lambda^2-|\lambda|\ \|\omega_n\|-\|(A+kQ)\varphi_n\|\cdot\|\omega_n\|.$$

We have $\lim\limits_{n\to\infty}\|\omega_n\|=0$, and from $(A-kQ)\varphi_n=\lambda\varphi_n+\omega_n$ obviously $\|(A-kQ)\varphi_n\|\leqq C$, where C does not depend on n. Hence $\|A\varphi_n\|\leqq|k|\ \|Q\varphi_n\|+C\leqq|k|\ \|A\varphi_n\|+C$,

$$\|A\varphi_n\|\leqq C/(1-|k|),\ \|(A+kQ)\varphi_n\|\leqq\|A\varphi_n\|+|k|\ \|A\varphi_n\|\leqq\frac{1+|k|}{1-|k|}C.$$

If $\lambda\neq 0$ it is therefore possible to choose a number $\boldsymbol{N}$ such that

$$-|\lambda|\ \|\omega_n\|-\|(A+kQ)\varphi_n\|\ \|\omega_n\|\geqq-\lambda^2/2 \qquad \text{for } n>\boldsymbol{N}.$$

For these n we have $2\lambda(A\varphi_n,\varphi_n)\geqq\lambda^2/2$, and from $(A\varphi_n,\varphi_n)\geqq 0$ it follows that $\lambda>0$. At any rate $\lambda\geqq 0$. No point λ of the spectrum of $A-kQ$ is negative. This means $((A-kQ)u,u)\geqq 0$ for all u of $\boldsymbol{A}$. This is true for k with $-1<k<1$. The limit $k\to 1$ resp. $k\to -1$ yields $((A-Q)u,u)\geqq 0$ resp. $((A+Q)u,u)\geqq 0$, which is the affirmed inequality $|(u,Qu)|\geqq(u,Au)$.

[7] I owe this to F. A. Ficken.

Heinz's inequality is nonelementary in the sense that it is no longer true if the assumption "A essentially self-adjoint in $\boldsymbol{D}$" is replaced by the weaker assumption "A Hermitian in $\boldsymbol{D}$." This is shown by the following example. We choose as Hilbert space the set of functions $u(x)$ defined in $0<x<\infty$ with finite

$$\left(\int_0^\infty |u|^2 dx\right)^{\frac{1}{2}}=||u||, \qquad (v, u)=\int_0^\infty \overline{v(x)}u(x)dx$$

and as $\boldsymbol{D}$ the subspace of all $u(x)$ with continuous second derivatives in $0<x<\infty$ vanishing identically in neighborhoods of $x=0$ and $x=\infty$. We define $Au=-u''$ and $Qu=\frac{3}{4}x^{-2}u$ for u in $\boldsymbol{D}$. The operator A, as well as Q, is Hermitian in $\boldsymbol{D}$. The operator Q in $\boldsymbol{D}$ is even essentially self-adjoint with $Q\geqq 0$ and the closure of A in $\boldsymbol{D}$ is an Hermitian operator for which v. Neumann's deficiency-indices are 1, 1.

By partial integration and Schwarz's inequality, one proves

$$\frac{9}{16}\int_0^\infty x^{-4}|u|^2 dx \leqq \int_0^\infty |u''|^2 dx;$$

hence $(Qu, Qu)\leqq(Au, Au)$.

The identity

$$\int_0^\infty |u'|^2 dx-\frac{1}{4}\int_0^\infty x^{-2}|u|^2 dx=\int_0^\infty \left|u'-\frac{1}{2x}u\right|^2 dx,$$

which holds for u in $\boldsymbol{D}$ shows $(u, Au)\geqq 1/3(u, Qu)\geqq 0$ for u of $\boldsymbol{D}$. In order to get a contradiction to (3) we only have to find a function u of $\boldsymbol{D}$ for which $(u, Qu)>0$ and $1>(u, Au)/(u, Qu)$. Define a sequence of functions $u_n(x)$ (which are not functions of $\boldsymbol{D}$) by

$$u_n(x)=\begin{cases} 0, & 0\leqq x\leqq 1/2n \\ (x-1/2n)2\sqrt{n}, & 1/2n\leqq x\leqq 1/n \\ x^{1/2}, & 1/n\leqq x\leqq n \\ (2n-x)/\sqrt{n}, & n\leqq x\leqq 2n \\ 0, & 2n\leqq x\leqq \infty. \end{cases}$$

For these functions compute

$$T(u_n)=\int_0^\infty |u_n'|^2 dx\Big/\int_0^\infty x^{-2}|u_n|^2 dx=\frac{\displaystyle\int_{1/n}^n \frac{dx}{4x}+O(1)}{\displaystyle\int_{1/n}^n \frac{dx}{x}+O(1)} \quad \text{for} \quad n\to\infty.$$

Thus $\lim\limits_{n\to\infty} T(u_n)=\frac{1}{4}$, and one can choose a number n_0 for which $T(u_{n_0})<\frac{1}{3}$. It is easy to find a function $u\neq 0$ of $\boldsymbol{D}$ for which $|T(u_{n_0})-T(u)|<\epsilon$ with prescribed $\epsilon>0$. We choose a u of $\boldsymbol{D}$ for which $T(u)<\frac{1}{2}$. From

$$T(u)=\frac{3}{4}\frac{(u, Au)}{(u, Qu)}$$

we have $\frac{1}{2}>\frac{3}{4}(u, Au)/(u, Qu)$, $1>(u, Au)/(u, Qu)$.

14. Bounds for Characteristic Roots of Matrices

Alfred Brauer[1]

For many applications in different fields it is of importance to determine the characteristic roots of a matrix approximately or to find at least small regions that contain the characteristic roots. Theoretically, this is a very simple problem. Because the characteristic equation of a matrix of order n is an algebraic equation of degree n, any method for the approximation of the roots of an algebraic equation could be used.

But in practice, this is only possible for very small values of n, since otherwise the computation of the coefficients of the characteristic equation is very tedious. Therefore, it is of importance to have methods that give approximations for the roots without having to compute the coefficients of the equation.

A large number of papers have been published on this subject. I do not want to give a historical account here. This is done in three papers by E. T. Browne [1], Olga Taussky [2], and W. V. Parker [3].

Let $A=(a_{\kappa\lambda})$ be an arbitrary square matrix of order n. We set

$$P_\kappa=\sum_{\substack{\lambda=1\\ \lambda\neq\kappa}}^{n}|a_{\kappa\lambda}| \qquad (\kappa=1,2,\ldots,n). \tag{1}$$

It is well known (see, for instance, [4]) that each characteristic root of A lies in the interior or on the boundary of at least one of the n circles

$$|z-a_{\kappa\kappa}|\leqq P_\kappa \qquad (\kappa=1,2,\ldots,n). \tag{2}$$

A few years ago, I improved this result as follows [5].

Theorem 1. Each characteristic root of A must lie in the interior or on the boundary of at least one of the $n(n-1)/2$ ovals of Cassini

$$|z-a_{\kappa\kappa}||z-a_{\lambda\lambda}|\leqq P_\kappa P_\lambda \qquad (\kappa,\lambda=1,2,\ldots,n;\ \kappa\neq\lambda). \tag{3}$$

Every point of these ovals lies in the interior or on the boundary of at least one of the circles (2).

I want to improve theorem 1 a little further. The formulation of the following theorem is somewhat more difficult, but its application is as simple as that of theorem 1.

Theorem 2 [6]. *Let $A=(a_{\kappa\lambda})$ be a square matrix of order n and P_κ be defined by* (1). *Set*

$$P_{\kappa\lambda}=|a_{\kappa\lambda}|P_\lambda+|a_{\lambda\kappa}|(P_\kappa-|a_{\kappa\lambda}|)+\sum_{\nu}|a_{\kappa\nu}a_{\lambda\nu}|+\sum_{\nu<\mu}|a_{\kappa\nu}a_{\lambda\mu}+a_{\kappa\mu}a_{\lambda\nu}|, \tag{4}$$

where κ, λ, μ, and ν run from 1 to n, and where $\kappa\neq\lambda$; $\mu\neq\kappa$, λ; $\nu\neq\kappa$, λ. Then each characteristic root ω of A lies in the interior or on the boundary of at least one of the n $(n-1)/2$ ovals of Cassini

$$|z-a_{\kappa\kappa}||z-a_{\lambda\lambda}|\leqq P_{\kappa\lambda} \quad (\kappa,\ \lambda=1,2,\ldots,n;\ \kappa\neq\lambda). \tag{5}$$

It is obvious that

$$P_{\kappa\lambda}\leqq P_\kappa P_\lambda. \tag{6}$$

We have the equality sign in (6) *if, and only if,*

$$\sum_{\nu<\mu}|a_{\kappa\nu}a_{\lambda\mu}+a_{\kappa\mu}a_{\lambda\nu}|=\sum_{\nu<\mu}|a_{\kappa\nu}a_{\lambda\mu}|+\sum_{\nu<\mu}|a_{\kappa\mu}a_{\lambda\nu}|.$$

[1] University of North Carolina.

Proof of theorem 2. Since ω is a characteristic root, the system of linear equations

$$\sum_{\lambda=1}^{n} a_{\kappa\lambda}x_\lambda=\omega x_\kappa \qquad (\kappa=1,2,\ldots,n) \tag{7}$$

has a nontrivial solution $x_1, x_2, \ldots, x_n$. Assume that

$$|x_r|\geqq|x_s|\geqq\max|x_\nu| \qquad (\nu=1,2,\ldots,n;\quad \nu\neq r,s) \tag{8}$$

We consider the rth and the sth of the equations (7)

$$\omega x_r-a_{rr}x_r=\sum_{\substack{\nu=1\\ \nu\neq r}}^{n} a_{r\nu}x_\nu, \tag{9}$$

$$\omega x_s-a_{ss}x_s=\sum_{\substack{\nu=1\\ \nu\neq s}}^{n} a_{s\nu}x_\nu. \tag{10}$$

If $x_s=0$, then $x_\nu=0$ for every $\nu\neq r$. It follows from (9) that $\omega=a_{rr}$ since $x_r\neq 0$. This proves theorem 2 if $x_s=0$.

Assume now that $x_s\neq 0$. Multiplying the equations (9) and (10) we obtain

$$\begin{aligned}(\omega-a_{rr})(\omega-a_{ss})\,x_rx_s&=\Big(\sum_{\nu\neq r}a_{r\nu}x_\nu\Big)\Big(\sum_{\mu\neq s}a_{s\mu}x_\mu\Big)\\ &=a_{rs}x_s\sum_{\mu\neq s}a_{s\mu}x_\mu+a_{sr}x_r\sum_{\nu\neq r,s}a_{r\nu}x_\nu+\sum_{\eta\neq r,s}a_{r\nu}a_{s\nu}x_\nu^2+\sum_{\substack{\nu<\mu\\ \nu,\mu\neq r,s}}(a_{r\nu}a_{s\mu}+a_{r\mu}a_{s\nu})\,x_\nu x_\mu\end{aligned}$$

hence by (1), (4), and (8)

$$|\omega-a_{rr}||\omega-a_{ss}|\leqq|a_{rs}|P_s+|a_{sr}|(P_r-|a_{rs}|)+\sum_{\nu\neq r,s}|a_{r\nu}a_{s\nu}|+\sum_{\substack{\nu<\mu\\ \nu,\mu\neq r,s}}|a_{r\nu}a_{s\mu}+a_{r\mu}a_{s\nu}|=P_{rs}.$$

This proves theorem 2.

All the results that I obtained in earlier papers [5, 7] for the ovals (3) can be extended to the ovals (5). Only some of these theorems will be mentioned.

Theorem 3. Let $f_1(y), f_2(y), \ldots, f_n(y)$ be arbitrary polynomials and A a square matrix of order n. Let A^* be the matrix which has as νth column the elements of the νth column of the matrix $f_\nu(A)$ for $\nu=1, 2, \ldots, n$. Denote the elements of A^* by $a^*_{\kappa\lambda}$ and the sums corresponding to the sums (1) and (4) by P^*_κ and $P^*_{\kappa\lambda}$. Then each characteristic root ω of A satisfies at least one of the $n(n-1)/2$ inequalities

$$|f_\kappa(\omega)-a^*_{\kappa\kappa}||f_\lambda(\omega)-a^*_{\lambda\lambda}|\leqq P^*_{\kappa\lambda}\leqq P^*_\kappa P^*_\lambda \qquad (\kappa,\lambda=1,2,\ldots,n;\ \kappa\neq\lambda).$$

If we use this theorem (see [7]) for the matrix

$$\begin{pmatrix}9&0&0&1&1\\1&2&2&1&0\\0&1&3&2&0\\0&1&1&1&0\\1&1&1&2&1\end{pmatrix}$$

and take for the polynomials $f_\nu(y)$ suitable cubic polynomials, then we obtain for the absolutely greatest characteristic root ω the inequality $9.1686<\omega<9.1851$ while actually $9.17267<\omega<9.17268$. Here the error for the lower bound is less than 0.05 percent.

A matrix with a dominant main diagonal is a matrix for which

$$|a_{\kappa\kappa}|>P_\kappa \qquad (\kappa=1,2,\ldots,n). \tag{11}$$

Minkowski [8] has already considered such matrices in the special case in which all the main diagonal elements are positive, and all the other elements are negative. He proved that their determinants are positive. The same holds for the determinants of all matrices which satisfy (11). This generalization of Minkowski's theorem follows at once from the fact that the characteristic roots lie in the interior or on the boundary of the circles (2). These circles do not intersect the imaginary axis if (11) holds. Hence no characteristic root equals 0, and the determinant, the product of the characteristic roots, cannot vanish.

Similarly, it follows from theorem 1 that the determinant cannot vanish if

$$|a_{\kappa\kappa}a_{\lambda\lambda}|>P_\kappa P_\lambda \qquad (\kappa,\ \lambda=1,2,\ldots,n;\quad k\neq\lambda) \tag{12}$$

holds instead of (11). When I published this theorem [5], I did not know a theorem of Ostrowski [9] from which the same result follows without using the characteristic roots. If Ostrowski's theorem is applied to the characteristic determinant of an arbitrary square matrix, then theorem 1 can be obtained. The inequalities (12) imply that (11) holds for all but one κ.

We now obtain from theorem 2

Theorem 4 [6]. *Assume that*

$$|a_{\kappa\kappa}a_{\lambda\lambda}|>P_{\kappa\lambda} \qquad (\kappa,\lambda=1,2,\ldots,n;\ \kappa\neq\lambda). \tag{13}$$

Then the determinant of A does not vanish. If moreover $a_{kk}>0$ for $\kappa=1,2,\ldots,n$, and if the characteristic equation has real coefficients, then the determinant of A is positive.

Further, theorem 4 generalizes Minkowski's theorem. It follows, for instance, that the determinant

$$D=\begin{vmatrix} a & 1 & 1 & 1 & 1\\ -1 & a & 1 & -1 & -1\\ 1 & -1 & a & 1 & -1\\ 1 & -1 & 1 & a & 1\\ -1 & -1 & 1 & 1 & a\end{vmatrix}>0$$

for $a>12^{\frac{1}{2}}$ since the inequalities (13) are satisfied. If moreover $a<4$, then each element of the main diagonal is smaller than the sum of the absolute values of the other elements of the same row.

Denote by $\alpha_1, \alpha_2, \ldots, \alpha_k$ any k different integers of the set $1, 2, \ldots, n$ and by $\beta_1, \beta_2, \ldots, \beta_{n-k}$ the remaining $n-k$ integers of this set. I call a square matrix of order n reduced if it is possible to find a set of integers $\alpha_1, \alpha_2, \ldots, \alpha_k$ such that all the elements of the matrix vanish which are in common to the rows $\alpha_1, \alpha_2, \ldots, \alpha_k$ and the columns $\beta_1, \beta_2, \ldots, \beta_{n-k}$. Otherwise I call the matrix unreduced.

G. Frobenius [10] uses the words "zerfallend" or "zerlegbar" for such matrices; V. Romanovsky [11] the word "décomposable."

O. Taussky [12] proved that a characteristic root of an unreduced matrix either lies in the interior of at least one of the circles (2) or on the boundary of each of them.

This theorem can be extended as follows:

Theorem 5 [13]. *A characteristic root of an unreduced matrix either lies in the interior of at least one of the $n(n-1)/2$ ovals* (5) *or on the boundary of each of them.*

Using the fact that the roots of an algebraic equation change continuously if the coefficients are changed continuously, we can prove the following result.

For each k we consider the $n-1$ ovals

$$|z-a_{kk}||z-a_{\lambda\lambda}|\leqq P_{k\lambda} \qquad (\lambda=1,2,\ldots,n;\ \lambda\neq k) \tag{14}$$

and take of each of these ovals only that simply connected region bounded by (14) which contains the point a_{kk}. We denote the closed region formed by these n simply connected regions by H_k.

Theorem 6 [6]. *If one of the regions H_k has no point in common with the region formed by all the other H_λ, then H_k contains one and only one characteristic root of A.*

A square matrix $A=(a_{\kappa\lambda})$ of order n is called stochastic if all the elements are nonnegative and if

$$\sum_{\lambda=1}^{n} a_{\kappa\lambda}=1 \qquad (\kappa=1,2,\ldots n). \tag{15}$$

The properties of the characteristic roots of stochastic matrices are of importance in the theory of stochastic processes.

It is well known that all the characteristic roots of a stochastic matrix lie in the interior or on the boundary of the unit circle. The point $z=1$ is always a characteristic root. R. v. Mises [20] pointed out that the results of G. Frobenius [14, 15, 10] on matrices with positive and nonnegative elements can be used for stochastic matrices and V. Romanovsky [11] formulated these results for stochastic matrices.

In particular, the following theorem holds. If A is unreduced, then $z=1$ is a simple characteristic root. No other point on the boundary of the unit circle can be a characteristic root unless all the elements of the main diagonal vanish.

N. Dmitriev and E. Dynkin [16] proved that no characteristic root of a stochastic matrix of order less than or equal to n can lie in the interior of the segments bounded by the unit circle and the chords joining the point $z=1$ with the points $z=e^{2\pi i/n}$ and $z=e^{-2\pi i/n}$. In a second paper [17], they generalized this result.

Let a_{ii} be the smallest element of the main diagonal of a stochastic matrix. M. Fréchet [18, 19] proved that all the characteristic roots lie in the interior or on the boundary of the circle

$$|z-a_{ii}|\leqq 1-a_{ii}. \tag{16}$$

Moreover, without using the results of Frobenius, he proved that no point different from 1 on the boundary of the unit circle can be a characteristic root unless at least two of the elements of the main diagonal vanish.

Using theorem 1 we can improve Fréchet's theorem as follows:

Theorem 7 [13]. *If a_{ii} and a_{jj} are the smallest elements of the main diagonal of a stochastic matrix, then all the characteristic roots lie in the interior or on the boundary of the oval of Cassini*

$$|z-a_{ii}|\,|z-a_{jj}|\leqq(1-a_{ii})(1-a_{jj}). \tag{17}$$

The proof can be obtained by elementary geometry by showing that each point of the ovals (3) lies in the interior or on the boundary of the oval (17).

If $a_{ii}\neq a_{jj}$, then the oval (17) lies in the interior of the circle (16). The boundaries of both curves have only the point $z=1$ in common. Hence (17) gives a better bound for the nontrivial characteristic roots, that is, the roots different from 1.

Often this result can be improved further.

Let ω and η be two different characteristic roots of an arbitrary square matrix A, and

$$\boldsymbol{x}=(x_1,\ x_2,\ \ldots,\ x_n)$$

a characteristic vector belonging to ω with regard to the rows of A and $\boldsymbol{y}$ a characteristic vector belonging to η with regard to the columns of A. It follows from E. Schmidt's theory of linear integral equations that $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal. This can easily be proved independently. We only use it for the special case that A is stochastic and $\omega=1$.

Since $\boldsymbol{x}=(1,\ 1,\ \ldots,\ 1)$ is a characteristic vector belonging to $\omega=1$ with regard to the rows, we have to prove the following theorem.

Theorem 8 [13]. *Let A be a stochastic matrix. If η is a nontrivial characteristic root and*

$$\boldsymbol{y}=(y_1,\ y_2,\ \ldots,\ y_n)$$

a characteristic vector belonging to η with regard to the columns of A, then

$$y_1+y_2+\ \ldots\ +y_n=0. \tag{18}$$

Proof. We have

$$\eta y_\lambda=\sum_{\kappa=1}^{n} a_{\kappa\lambda}y_\kappa \qquad (\lambda=1,2,\ldots,n). \tag{19}$$

If we add these equations, then we obtain by (15)

$$\eta(y_1+y_2+\ .\ .\ .\ +y_n)=y_1+y_2+\ .\ .\ .\ +y_n.$$

This proves (18) since $\eta=1$.

If a matrix A is reduced, then its characteristic determinant is the product of characteristic determinants of unreduced matrices and the characteristic roots of A are the characteristic roots of these unreduced matrices. Hence it is sufficient if we consider unreduced matrices.

Theorem 9 [13]. *Let $A=(a_{\kappa\lambda})$ be an unreduced stochastic matrix and $h_1, h_2, \ldots, h_n$ arbitrary numbers. Denote the matrix $(a_{\kappa\lambda}-h_\lambda)$ by B. Then B has the trivial characteristic root*

$$\omega'=1-\sum_{\lambda=1}^{n} h_\lambda, \tag{20}$$

and the other roots are the nontrivial characteristic roots of A with the respective multiplicities.

Proof. It follows from (15) and (20) that

$$\sum_{\lambda=1}^{n}(a_{\kappa\lambda}-h_\lambda)=1-\sum_{\lambda=1}^{n}h_\lambda=\omega';$$

hence ω' is a characteristic root of B and $(1, 1, \ldots, 1)$ is a characteristic vector belonging to ω' with regard to the rows. Since A is unreduced, $\omega=1$ is a simple root.

Let η be another root of A belonging to the vector $\boldsymbol{y}=(y_1, y_2, \ldots, y_n)$ with regard to the columns. It follows from (19) and (18) that

$$\sum_{\kappa=1}^{n}(a_{\kappa\lambda}-h_\lambda)y_\kappa=\sum_{\kappa=1}^{n}a_{\kappa\lambda}y_\kappa-h_\lambda\sum_{\kappa=1}^{n}y_\kappa=\sum_{\kappa=1}^{n}a_{\kappa\lambda}y_\kappa=\eta y_\lambda.$$

Hence η is a characteristic root of B belonging to $\boldsymbol{y}$ with regard to the columns. It follows in the same way that each nontrivial characteristic root of B is a characteristic root of A.

Since the roots of the characteristic equation change continuously if the elements of the matrix are changed continuously, it follows that the corresponding roots of A and B have the same multiplicities unless ω' is a characteristic root of A. In this case the multiplicity of ω' in B is greater by one.

It follows from theorem 8 that we can apply all the theorems on bounds for the characteristic roots to B in order to obtain bounds for the nontrivial characteristic roots of A. This method is especially efficient if for each m, $\max|a_{\rho m}-a_{\sigma m}|$ $(\rho, \sigma=1, 2, \ldots n;\ \rho, \sigma\neq m)$ is relatively small. Let us consider, for instance

$$A=\begin{pmatrix} .33 & .27 & .21 & .19 \\ .42 & .14 & .23 & .21 \\ .44 & .26 & .12 & .18 \\ .42 & .27 & .22 & .09 \end{pmatrix}$$

We choose $h_1=.42$, $h_2=.27$, $h_3=.22$, $h_4=.19$ and obtain

$$B=\begin{pmatrix} -.09 & 0 & -.01 & 0 \\ 0 & -.13 & .01 & .02 \\ .02 & -.01 & -.10 & -.01 \\ 0 & 0 & 0 & -.10 \end{pmatrix}$$

The trivial characteristic root of B is $-.1$; the other characteristic roots are the roots of the principal minor B_{44} of order 3 of B. It follows from theorems 1, 5, and 6 applied to B_{44} that one nontrivial characteristic root of A must lie in the interval $-.14<z<-.12$ and the other two in the interior of at least one of the ovals $|z+.09||z+.10|\leqq .0003, |z+.10||z+.13|\leqq .0003$.

The following special case is of interest:

Theorem 9. If b is one of the elements of the main diagonal of a stochastic matrix while all the other elements of the column of b equal c, then b-c is a characteristic root.

[1] E. T. Browne, Limits to the characteristics roots of a matrix, Am. Math. Monthly **46,** 252–265 (1939).

[2] O. Taussky, A recurring theorem on determinants, Am. Math. Monthly **56,** 672–676 (1949).

[3] W. V. Parker, Characteristic roots and field of values of a matrix, Bull. Am. Math. Soc. **57,** 103–108 (1951).

[4] A. Brauer, Limits for the characteristic roots of a matrix, Duke Math. J. **13,** 387–395 (1946).

[5] A. Brauer, Limits for the characteristic roots of a matrix, II, Duke Math. J. **14,** 21–26 (1947).

[6] A. Brauer, Limits for the characteristic roots of a matrix, V. Duke Math. J. **19,** 553–562 (1952).

[7] A. Brauer, Limits for the characteristic roots of a matrix, III, Duke Math. J. **15,** 871–877 (1948).

[8] H. Minkowski, Zur Theorie der Einheiten in algebraischen Zahlkörpern, Göttinger Nachr., pp. 90–93 (1900).

[9] A. Ostrowski, Ueber Determinanten mit überwiegender Hauptdiagonale, Comm. Math. Helv. **10,** 69–96 (1937).

[10] G. Frobenius, Ueber Matrizen aus nicht negativen Elementen, Sitzungsber. Berlin, pp. 456–477 (1912).

[11] V. Romanovsky, Recherches sur les chaines de Markoff, Acta Math 66, 147–251 (1936).

[12] O. Taussky, Bounds for characteristic roots of matrices, Duke Math. J. **15,** 1043–1044 (1948).

[13] A. Brauer, Limits for the characteristic roots of a matrix, IV, Applications to stochastic matrices, Duke Math. J. **19,** 75–91 (1952).

[14] G. Frobenius, Ueber Matrizen aus positiven Elementen, Sitzungsber. Berlin, pp. 471–476 (1908).

[15] G. Frobenius, Ueber Matrizen aus positiven Elementen, II, Sitzungsber. Berlin, pp. 514–518 (1909).

[16] N. Dmitriev and E. Dynkin, On the characteristic roots of a stochastic matrix, Comptes rendus acad. URSS **49,** 159–162 (1945).

[17] N. Dmitriev and E. Dynkin, On the characteristic roots of stochasic matrices, Bull. acad. sci. URSS, Ser. Math. **10,** 167–184 (1946).

[18] M. Fréchet, Comportement asymptotique des solutions d'un système d'équations linéaires et homogènes aux différences finies du premier order à coefficients constants, Pub. faculté sci. univ. Masaryk **178,** 1–24 (1933).

[19] M. Fréchet, Recherches théoriques modernes sur la théorie des probabilités, pt. 2 (E. Borel, Traité du calcul des probabilités et ses applications I, 3) (Gauthier-Villars, Paris, 1938).

[20] R. v. Mises, Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik (Leipzig, Deuticke, 1931).

15. Matrix Inversion and Solution of Simultaneous Linear Algebraic Equations With the IBM 604 Electronic Calculating Punch

George W. Petrie, III[1]

Introduction

Current computational literature is replete with many descriptions of methods for matrix inversion. (See references.) Gutshall [1] points out the necessity for further statistical study of types of matrices which may be subjected to numerical inversion, and also suggests a comparative study of known methods of inversion. Dantzig [2] demonstrates in his simplex method a recurring need for obtaining the inverse of certain matrices to compute optimum programs. In fact it is his request for a simple fast procedure for inversion of large order Leontief type matrices that led to the technique presented in this paper. The procedure, while using the standard elimination method [3, 4, 5, 6, 7, 8] allows of simple and continuous processing of cards through a reproducer, the 604, an accounting machine (for checking), and a second reproducer. This cycle is repeated N times to invert an Nth order matrix. The size of the matrix is not limited and the four machines are operated continuously without change of control panels.

Consider the matrix equation, $AX=I$, as describing a system of simultaneous linear equations in the x's. By simple operations (multiplying equations by appropriate constants and combining with other equations) it is possible (provided A is nonsingular) to reduce the system to one in which the coefficient of X is I. In this case, the right hand member of the matrix equation is reduced to the inverse of A. Thus, in the elimination method, one starts with an augmented matrix composed of the original matrix, A, and the unit matrix, I. As the original matrix is reduced to the unit matrix, the original unit matrix is simultaneously reduced to the inverse of the original given matrix. This reduction may be carried out one column at a time. After $(k-1)$ columns have been reduced, the procedure involves dividing the kth row by a_{kk} (to obtain 1 in the kth column) and then subtracting the product of this result with the appropriate constant, a_{ik} from the ith row so as to obtain 0 in the kth column of the ith row. This latter subtraction is carried out for all $i \neq k$. Columns of the unit matrix numbered greater than k are unaffected. The kth column is, in general, completely changed. Thus N^2 elements enter the computation at any one cycle. In the present procedure, the kth column (or vector) of the left hand part of the reduced matrix is not written since its value is known automatically. Instead, it is replaced by the kth column (or vector) of the right hand part of the augmented matrix after this step in the reduction. By selecting the appropriate formula to use (as listed in step 3 of the next section) the 604 actually brings in vectors from the unit matrix as needed. The same machine is used to allow the matrix elements to be reordered after each step in the elimination so that the new pivotal row is on top and the new pivotal column is at the left. Thus intermediate row and column sorting is entirely eliminated. It is through the elimination of all collating and this type of sorting that the present more rapid machine procedure becomes possible.

The accounting machine is used for checking purposes. The checking operation must be looked upon as occupying one step in the continuous flow of cards through the four machines. Processing would not be appreciably accelerated if this check were omitted since all machines are running simultaneously. It is important to point out, however, that the checking procedure augments the original N^2 cards to (N^2+N) cards.

Machine Procedure

The original Nth order matrix $((a_{ij}))$ is punched on N^2 cards, one element, together with the identifying i and j, to a card. An additional row of check sums is appended to the original matrix. These

[1] International Business Machines, Washington, D. C.

values are defined by $s_j=+1-\sum_i a_{ij}$. A 10-digit fixed decimal system is used (two integers and eight decimals). Data is punched as follows:

Columns:	*Data*
1 to 3	Number of row (i)
4 to 6	Number of column (j)
10 to 19	Value of element (a_{ij})
20 to 29	Value of element (a_{i1})—first card only.

In all cases minus signs are carried as x-punches over the right-hand digit. In addition, elements of the first row are identified with x and y punches in column 1 (denoted $xy1$) to indicate the pivotal row; elements of the first column x and $y7$ punches to indicate the pivotal column. Elements of the row of s_j are identified by x and $y13$ punches. The 521 unit is used to supply subsequent x and y punches as needed.

After the cards are prepared and sorted to row within column, the first column is gang-punched as listed in step 6. After this preliminary step, the following six steps are repeated N times to obtain the inverse:

1. Reproduce the values of i (columns 1 to 3) and a_{ij} (columns 10 to 19) from the leading $(N+1)$ cards (with identifying $xy7$) into each $N+1$ of the remaining $(N-1)$ $(N+1)$ cards.

Read ($xy7$)	*Punch* ($nxy7$)
Digits of 1 to 3	1 to 3
10 to 19	20 to 29

This means that an element in the ith row of the leading column will now appear on the same card with all other elements of the ith row of the matrix. As a variation, the value of a_{ij} to be reproduced may be read from either columns 10 to 19 of $xy7$ cards or from columns 20 to 29 of $nxy7$ (no $xy7$) cards which have already passed through the punch side of the reproducer for this step. Thus, in practice, the operator starts with the $N+1$ cards bearing the designator $xy7$ in the read feed and all others either in the punch feed or immediately available thereto. After reproducing the values from the $N+1$ cards, both stackers are emptied. The $2(N+1)$ cards are placed together in the read feed and $4(N+1)$ cards are next obtained from the two stackers. As soon as a sizeable group of cards are generated, one may take the cards from the read stacker to the next operation process at the 521–604 while continuing to reproduce values from the cards that came from the punch stacker of the reproducer.

In both cases a comparison check is carried on all reproduction. In addition, the reproduced value of i (columns 1 to 3) is checked for double punches. The reason for the reproduction of i is apparent at step 3.

2. Place a blank card having a different distinguishing color and $xy1$ punch after the last card processed in step 1.

3. Calculate values of b_{ij} and t_j according to the following formulas. For complete generality, assume that $(k-1)$ steps in the inversion process have been completed, and that the matrix is in order of row within column as given in the following array:

$$\begin{array}{llllllll}
a_{k,k} & \cdots & a_{k,j} & \cdots & a_{k,N} & a_{k,1} & \cdots & a_{k,k-1} \\
\vdots & & \vdots & & & & & \vdots \\
a_{i,k} & \cdots & a_{i,j} & & & & & \\
\vdots & & & & & & & \\
a_{N,k} & & & & & & & \\
s_k & \cdots & s_j & & & & \cdots & s_{k-1} \\
a_{1,k} & & & & & & & \\
\vdots & & & & & & & \vdots \\
a_{k-1,k} & & & & & & \cdots & a_{k-1,k-1}
\end{array}$$

alues of b_{ij} and $t_j = +1 - \sum_i b_{ij}$ are computed by:

	xy—conditions
$b_{k,k} = \frac{1}{a_{k,k}}$	pivotal element $xy1$, $xy7$
$b_{i,k} = -\frac{1}{a_{k,k}} \cdot a_{i,k}$ for $i \neq k$	pivotal column $nxy1$, $xy7$
$b_{k,j} = +\frac{a_{k,j}}{a_{k,k}}$ for $j \neq k$	pivotal row $xy1$, $nxy7$
$b_{i,j} = a_{i,j} - \frac{a_{k,j}}{a_{k,k}} \cdot a_{i,k}$ for $i \neq k, j \neq k$	$nxy1$, $nxy7$
$t_k = -\frac{1}{a_{k,k}} \cdot s_k$	check, pivotal column $nxy1$, $xy7$
$t_j = s_j - \frac{a_{k,j}}{a_{k,k}} \cdot s_k$ for $i \neq k$	check, other $nxy1$, $nxy7$

ne observes that the computation for t_j is identical with that for $b_{i,j}$ where $i \neq k$. The values computed are punched directly into columns 30–39 on all $nxy1$ cards. For $xy1$ cards, however, the values are tored and punched on the following $xy1$ card. This results in the following array of values:

$$
\begin{array}{cccccccccc}
 & b_{k,k} & \cdots & b_{k,j-1} & \cdots & b_{k,N-1} & b_{k,N} & \cdots & b_{k,k-2} & b_{k,k-1} \\
b_{k+1,k} & b_{k+1,k+1} & \cdots & b_{k+1,j} & \cdots & b_{k+1,N} & b_{k+1,1} & \cdots & b_{k+1,k-1} & \\
\vdots & & & \vdots & & \vdots & \vdots & & \vdots & \\
b_{i,k} & & & b_{i,j} & \cdots & b_{i,N} & b_{i,1} & \cdots & b_{i,k-1} & \\
\vdots & & & \vdots & & & & & \vdots & \\
b_{N,k} & & & & & & & & & \\
t_k & & \cdots & t_j & & & & \cdots & t_{k-1} & \\
b_{1,k} & & & & & & & & & \\
\vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots & \\
b_{k-1,k} & b_{k-1,k+1} & \cdots & b_{k-1,j} & \cdots & b_{k-1,N} & b_{k-1,1} & \cdots & b_{k-1,k-1} &
\end{array}
$$

Notice that no b is punched on the first card. The additional card of step 2 was added to obtain the ast value. In addition to the above calculations, the following operations are performed simultaneously n the 521:

(a) Offset gang punch digits of columns 4 to 6 into 7 to 9.

(b) Offset gang punch $xy1$ into $xy4$.

(c) Gang punch $xy7$ into $xy7$ on the $xy1$ card following an $xy7$ card.

(d) Emit $xy10$ into cards following the $xy1$ which follows an $xy7$. Stop the $xy10$ emission after the next $xy1$ card.

t will be observed that these operations result in a column identification of the last card (the card added t step 2). As yet there is no row identification, but this will be supplied by step 1 of the following cycle. f needed for a visual check, it is easy to remember that the unidentified row is one greater than the olumn number that is already on the card. It would also be possible to prepunch the row identification efore step 2. This would involve a simple consecutive number deck.

4. Tabulate and list N lines with the following data:

$$\begin{array}{l} k \quad \text{(read from first card of column)} \\[1em] \sum b_{i,j} \\[1em] t_j \\[1em] t_j+\sum_i b_{i,j}. \end{array}$$

This tabulation is used to check the accuracy of the previous steps. The last value should differ from $+1$ by no more than an acceptable rounding error. One has a choice of checking each column separately or else the entire matrix as a whole.

5. Remove the first card, of which no further use is made. Take the next $N+1$ cards, and place them at the end of the deck. This leaves the cards in the following array:

$$\begin{array}{cccccc}
b_{k+1,k+1} & \cdots & b_{k+1,N} & b_{k+1,1} & \cdots & b_{k+1,k} \\
\vdots & & \vdots & \vdots & & \vdots \\
b_{N,k+1} & & & & & \\
t_{k+1} & \cdots & t_n & t_1 & \cdots & t_k \\
b_{1,k+1} & & & & & \\
\vdots & & \vdots & \vdots & & \vdots \\
b_{k,k+1} & \cdots & b_{k,N} & b_{k,1} & \cdots & b_{k,k}
\end{array}$$

6. Reproduce into a new set of cards (or tumbled cards already half utilized).

Read	*Punch*
Digits of 1 to 3	1 to 3
Digits of 7 to 9	4 to 6
30 to 39	10 to 19 on all cards
$xy4$	$xy1$
$xy10$	$xy7$
$xy13$	$xy13$

At the same time gang punch from 10 to 19 of $xy1$, $xy7$ into 20 to 29 on all $nxy1$, $xy7$ cards.

The new set of cards will be in such order that the pivotal row is on top, the pivotal column to the left, and elements of both are marked with proper identifying x and y punches. Thus these cards are ready to process through step 1 of the following cycle. After N cycles the values obtained at step 5 give the required inverse.

Solution of Equations

To solve a set of linear simultaneous algebraic equations, the procedure outlined above is only slightly modified. The same control panels are used without any changes or additional wiring. The original matrix of coefficients is augmented by the vector of constants and row of negative sums of column elements for checking. Thus one starts with a matrix of $(N+1)^2$ elements. After step 1 on each cycle, the first $(N+1)$ cards ($xy7$) are discarded. Thus the cards which pass through the 604 number $N(N+1)$,

$(N-1)(N+1)$, $(N-2)(N+1)$ . . ., $1(N+1)$. At step 4 the check value is approximately -1, -2, . . . , $-N$ on the successive steps and should differ from these negative integers by no more than an acceptable rounding error. At step 5 the first card is removed. No other rearranging occurs. Finally, one obtains the solution of the simultaneous equations as a vector accompanied by a check sum. This procedure should take approximately half as long as the inversion procedure.

Summary

The inversion procedure outlined above is believed to be faster and easier to perform than other methods now in common practice. Further inquiry should be made of the applicability of this procedure to a floating decimal calculation, also to matrices involving complex numbers. The author understands that members of Dr. Dantzig's group are investigating the problem of which order of elements allows of the determination of the most accurate inverse. Certainly more work is necessary in this area.

For large-order matrices this procedure may be used to continuously process cards from one machine to another, keeping all machines in operation simultaneously. For small-order matrices, the number of cards is insufficient to keep all machines running. In this case, all cards would be processed at one machine at a time. Several matrices could be inverted at the same time.

It is possible to perform steps 1 and 6 on a single type 519 document originating machine with only one control panel if a control punch has been emitted onto the cards at step 3. This modification would be useful for handling low-order matrices.

If desired, it is also possible to handle a row check similar to the column check. In this case, the check sum is defined by:

$$s_i = -1 - \sum_j a_{ij}$$

The writer expresses his acknowledgments to Dr. George B. Dantzig, U. S. Air Force Comptroller's Office; Dr. Jack Sherman, The Texas Co.; and Miss Caroline J. Cooper, IBM. Dr. Dantzig requested the development of a fast method of inversion of Leontief type matrices. He also suggested the technique whereby the unit matrix is generated vector by vector as needed, resulting in the equations for $b_{i,j}$ in step 3. Dr. Sherman developed the reproduction used in step 1 in a paper concerning matrix inversion with the 602A.[9] The assistance of Miss Cooper was realized in aiding with the programming and executing the entire plan into successful operation.

References

[1] W. D. Gutshall, Practical inversion of matrices of high order, Computation Seminar, pp. 171–173 (IBM) (December 1949).

[2] G. B. Dantzig, Maximization of a linear function of variables subject to linear inequalities, Activity Analysis of Production and Allocation, chap. 21 (John Wiley & Sons, 1951).

[3] P. S. Dwyer, Linear Computations, pp. 190–191 (John Wiley & Sons, 1951).

[4] R. A. Frazer, W. J. Duncan, and A. R. Collar, Elementary matrices, pp. 119, 120 (Macmillan Publishing Co., New York, N. Y., 1947).

[5] K. S. Kunz, Matrix methods, Computation Seminar, pp. 37–42 (IBM) (December 1949).

[6] J. Lowe, Solution of simultaneous linear algebraic equations using the IBM type 604 electronic calculating punch, Computation Seminar, pp. 54–55 (IBM) (December 1949).

[7] W. E. Milne, Numerical calculus, pp. 15–29 (Princeton University Press, 1949).

[8] F. M. Verzuh, The solution of simultaneous linear equations with the aid of the 602 calculating punch, Mathematical Tables and Other Aids to Computation, III, pp. 453–462 (1949).

[9] J. Sherman, Computations of inverse matrices by means of IBM machines, Applied Science Department Technical Newsletter No. 3 (IBM).

ADDITIONAL REFERENCES

J. Chancellor, J. W. Sheldon, and G. Liston Tatum, The solution of simultaneous linear equations using the IBM card-programmed electronic calculator, Industrial Computation Seminar, pp. 57–61 (IBM) (September 1950).

H. H. Goldstine and J. von Neumann, Numerical inverting of matrices of high order, II, Proc. Am. Math. Soc. **2,** 188–202 (1951).

H. Hotelling, Some new methods in matrix calculation, Ann. Math. Stat. **14,** 1–34 (1943).

H. Hotelling, Further points on matrix calculation and simultaneous equations, Ann. Math. Stat. **14,** 440–441 (1943).

I. C. Liggett, The Gauss-Seidel method of solution of simultaneous linear equations, Industrial Computation Seminar, pp. 62–65 (IBM) (September 1950).

J. von Neumann and H. H. Goldstine, Numerical inverting of matrices of high order, Bul. Am. Math. Soc. **53,** 1021–1099 (1943).

16. Experiments on the Inversion of a 16X16 Matrix[1]

John Todd*

Introduction

Some experiments have been carried out in the Computation Laboratory of the National Bureau of Standards on the inversion of a certain 16×16 matrix, using the following three methods: (1) G. W. Petrie's arrangement of the Gauss elimination process,[2] (2) a Monte Carlo process,[3] and (3) an iteration method.[3] It is the purpose of this note to describe and compare the results obtained.

The matrix was a 16×16 matrix of the so-called Leontief type [1, 2],[4] representing certain inter-industry relations. A matrix of the same type, but of order 40×40, has been investigated by J. L. Holley of the Air Comptroller's Office, USAF; in particular, it has been inverted, using the UNIVAC.

The actual matrix inverted was $B=I-A$, where 10^5A is:

0	36265	0	0	0	0	0	0	4646	3169	8939	0	421	0	1355	0
5158	0	0	0	0	0	0	0	1382	103	2466	0	120	0	6995	5
188	0	0	9779	12795	2679	0	0	0	1062	0	0	0	909	0	5
20	38	77	0	30	128	145	264	29	85	99	335	0	0	5	0
69	2225	424	10425	0	128	0	275	882	874	83	0	601	2434	253	104
49	173	4165	2398	3252	0	145	550	2705	86	17	0	3065	0	0	30
128	913	1080	1209	696	319	0	66	3528	754	33	561	300	174	200	10
3794	1109	12225	1835	2087	3890	8475	0	5322	2861	1986	2578	2464	8653	2822	277
2915	905	1388	709	1361	638	1065	286	0	1867	5314	3475	2404	201	0	173
869	1682	77	709	907	383	1937	22	1764	0	761	448	1563	134	1194	6747
524	324	0	1501	121	64	97	0	382	1388	0	7960	1202	0	146	119
326	23	0	3878	242	0	0	11	29	34	463	0	60	348	58	139
10	0	0	250	136	0	0	0	0	137	1639	0	0	321	331	1288
7845	2942	10297	2252	1996	4719	14286	24210	6616	6630	844	3475	421	0	19	0
3537	2527	926	1877	1165	510	1840	385	3323	2895	3145	897	2764	642	0	515
5444	2821	501	1772	1119	191	630	429	5110	1079	1374	4596	962	976	5294	0

Method 1

The actual running time was some 8 hours on the 604 Calculating Punch. The error in the check sums carried was about 160 units in the last (eighth) place. The resultant matrix is denoted by G and is available.

*National Bureau of Standards.

[1] This work has been supported by the Air Comptroller's Office, USAF.

[2] See the previous paper. The IBM operations were carried out under the direction of Helen V. Hammar of the NBS Computation Laboratory.

[3] These were carried out on SEAC under the direction of Karl Goldberg of the NBS Computation Laboratory.

[4] See also [3] for description of the inversion of a 38×38 matrix of this type, by a Gaussian process, on the Aiken Relay Computer, Mk. II.

Method 2

The method described by G. E. Forsythe and R. A. Leibler [4] was used in the Monte Carlo experiments. The "random" numbers used were generated by the process suggested by Olga Taussky: residues modulus 2^{42} of powers of an odd power of 5. These residues have period 2^{40} and are generated by a single operation ("low" multiplication) on SEAC.

The first experiment involved carrying out 1,000 walks per row; this took some 21 minutes. The second used 10,000 walks per row and took 3 hours 15 minutes.

Reading in the (matrix and the) instructions took about 21 minutes; printing out the matrix took about 11 minutes.The complete matrices, denoted by M_0 and M, were printed out and are available.

Method 3

The iteration method used was the standard one [5, 6],

$$X_{n+1}=X_n(2-BX_n), \qquad n\geq 0.$$

The initial approximation actually used was $X_0=M_0$, that is, that obtained after the 1,000-walk-Monte Carlo experiment, but not very different results would have been obtained in the present case if we had taken $X_0=I$. Each iteration takes about a minute on SEAC. Printing out takes about 11 minutes as before, and about 15 minutes would be required for reading in the matrix and the instructions.

The following material is available: The original approximation $X_0=M_0$ and the following iterates X_i, $i=1(1)10,15,20$. In addition, for checks, the matrices, $R_{15}=BX_{15}-I$ and $R_{20}=BX_{20}-I$, were printed out.

The results have not yet been analyzed fully, but some idea of their accuracy can be obtained from the following table:

	Element (1,1)	Element (7,11)
$X_0=M_0$	1. 03528 3690	. 00348 1200
M_1	1. 02289 5670	. 00301 7210
X_5	1. 02259 5947	. 00322 1541
X_{10}	1. 02347 2122	. 00320 9509
X_{15}	1. 02349 5929	. 00320 8423
X_{20}	1. 02349 6765	. 00320 8375
G	1. 02349 680	. 00320 835

	R_{15}*	R_{20}*	$(G-X_{20})$**
Maximum element	117538	4726	387
RMS element	1137	44	25
(1,1)	−9703	−382	4
(7,11)	−568	23	3

*In units of ninth decimal.
**In units of eighth decimal (the elements of X_{20} being rounded)

Remarks

In the present case it would probably have been more efficient to omit the Monte Carlo calculation and start off with $X_0=I$.

Using current figures of cost there is little to choose between the estimates for the IBM method and the Monte Carlo—iteration method. However, if the Monte Carlo stage was omitted, and if faster input and output devices were available***, there is a considerable margin in favor of the SEAC operation.

The SEAC operation would handle larger matrices, say up to order 27, in the combined high speed memories. If still higher order matrices were to be handled use could be made of the magnetic tape.***

Further experiments on the inversion of the same matrix B and on the determination of its characteristic values are planned.

***(Added September 1952) Such devices are now in operation and current rates of input and output to magnetic wire are about 50 words per second. Using magnetic tapes for intermediate storage, operations on matrices of order 204 have been successfully carried out; the total capacity presently available is about 80,000 words.

References

[1] F. V. Waugh, Inversion of the Leontief matrix by power series, Econometrica **18,** 142–154 (1950).

[2] J. L. Holley, Note on the inversion of the Leontief matrix, Econometrica **19,** 317 to 320 (1951).

[3] H. F. Mitchell, Jr., Inversion of a matrix of order 38, Math. Tables and other Aids to Computation **3**, 161 to 166 (1948).

[4] G. E. Forsythe and R. A. Leibler, Matrix inversion by a Monte Carlo method, Math. Tables and other Aids to Computation **5,** 127 to 129 (1950).

[5] V. Bargmann, F. J. Murray, and J. von Neumann, Solution of linear systems of high order (Princeton, 1946).

[6] H. Hotelling, Some new methods in matrix calculation, Ann. Math. Stat. **14,** 1 to 34 (1943).

17. A Method of Computing Eigenvalues and Eigenvectors Suggested by Classical Results on Symmetric Matrices

Wallace Givens [1]

17.1. Introduction

The problem posed is precisely that considered by Goldstine in his lecture before the Symposium: to devise a method of calculating all the eigenvalues and eigenvectors of a real symmetric matrix suitable for use with an automatic sequenced high-speed digital computer. It can be regarded as a source of satisfaction that the basic technical device proposed here, that of a sequence of rotations in coordinate planes, is the same as that used by von Neumann and Goldstine. An essential difference is that the plane rotations are used only to reduce the data of the problem from $\frac{1}{2}n(n+1)$ numbers to $2n-1$, after which the roots of the matrix are obtained by operations with a Sturm sequence of polynomials. Thus for a matrix of order $n=100$, storage for 5,050 numbers is initially required, but in the first stages of the problem this data can be effectively processed in segments until only 199 numbers need be used when the eigenvalues are calculated. If the 10^4 components of the n eigenvectors are required, storage requirements must again grow large, but the interplay between the internal and external memories appears favorable. The advantages as to memory requirements are like those of the von Neumann-Goldstine method and differ from techniques that require the calculation of a sequence x, Ax, A^2x, . . ., of vectors and that involve operation with the whole matrix A at each step.

The fundamental problem of the effect of round-off error is not considered here. It seems likely that the conclusions stated by Goldstine will be applicable, but this needs further study.

Added in proof September 10, 1952: A detailed error analysis has now been made, and the stability of the proposed method of finding eigenvalues can be fully guaranteed.

17.2. Summary of Results

By a sequence of at most $\frac{1}{2}(n-1)(n-2)$ fully determined rotations in coordinate planes, a real symmetric matrix A of order n can be reduced to the form

$$S=\left\{\begin{array}{ccccccccc}
\alpha_1 & \beta_1 & 0 & 0 & 0 & . & . & . & 0 \\
\beta_1 & \alpha_2 & \beta_2 & 0 & 0 & . & . & . & 0 \\
0 & \beta_2 & \alpha_3 & \beta_3 & 0 & . & . & . & 0 \\
0 & 0 & \beta_3 & \alpha_4 & \beta_4 & . & . & . & 0 \\
. & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & \beta_{n-2} & 0 \\
. & . & . & . & . & . & \beta_{n-2} & \alpha_{n-1} & \beta_{n-1} \\
0 & 0 & 0 & 0 & 0 & . & 0 & \beta_{n-1} & \alpha_n
\end{array}\right\} \qquad (1)$$

[1] Oak Ridge National Laboratory and the University of Tennessee.

The successive principal minors $f_i(\lambda)$ of $\lambda I-S$ and the characteristic equation of $S, f_n(\lambda)=0$, (which is the same as that of A) are obtained by the recursion formula

$$f_i(\lambda)=(\lambda-\alpha_i)f_{i-1}(\lambda)-(\beta_{i-1})^2 f_{i-2}(\lambda), \tag{2}$$

where $f_{-1}=0, f_0=1$, and $i=1, 2, \ldots, n$.

If any $\beta_j=0$, S decomposes into the direct sum of two symmetric matrices, and the determination of eigenvalues and eigenvectors of S (and hence of A) can be replaced by the same problems for the two smaller matrices. For simplicity of statement, we assume $\beta_j \neq 0$ for $j=1, \ldots, n-1$, although "small" but nonzero, β_j should be dealt with more fully in setting up details of a computational procedure. Then $f_n(\lambda), f_{n-1}(\lambda), \ldots, f_1(\lambda), f_0=1$ form a Sturm sequence of functions. Hence the number of roots of $f_n(\lambda)$ greater than any chosen real number a, with $f_n(a)\neq 0$, is equal to the number of variations in sign of the sequence $f_n(a), f_{n-1}(a), \ldots, f_1(a), f_0(a)=1$. Moreover, the roots of each f_i are separated by those of f_{i-1} $(i=2, \ldots, n)$. The coefficients of $f_n(\lambda)$ can be computed if desired, but the recursion formula yields $f_n(a)$ for any a without much more computation than would be required if the coefficients were known. The knowledge of a Sturm sequence (without additional computation) makes it appear plausible that the roots of $f_n(\lambda)=0$ can now be computed to any required accuracy with as much ease as should be expected from the inherent difficulty of the problem.

Having found an eigenvalue ρ, an eigenvector x of S can be found by solving the equations $(\rho I-S)x=0$ for the ratios of the x_i; this is trivially done by the recursion formula

$$x_{i+1}=(\beta_i)^{-1}[(\rho-\alpha_i)x_i-\beta_{i-1}x_{i-1}],$$

where $i=1, \ldots, n-1$, $\beta_0=x_0=0$ and for $i=n$, $\beta_n=1$, and $x_{n+1}=0$ is a check on the computation. To improve the accuracy, an approximation to ρ could be used to find an approximate eigenvector and then both x and ρ simultaneously refined by any desired iterative process. Since only $2n-1$ numbers need to be stored to specify S, the chosen iterative process can be carried out for matrices of large order, using only the internal memory of a high-speed computing machine. To find the corresponding eigenvector of A from one of S, it is only necessary to "record" the reduction of A to S: if $S=T'AT$ with $T=T_{23}T_{24} \ldots T_{n-1n}$ (each T_{pq} being a plane rotation), so that T is obtained if the columns of the unit matrix are operated on precisely as are the columns of A, and $Sx=\rho x$, then $Ay=\rho y$ with $y=Tx$.

17.3. Heuristic Considerations

In a problem of the nature here considered, it seems certain that a wide variety of methods of solution will be needed because of varying requirements for accuracy, size of matrix involved, use of the method as a subroutine, etc. In particular, the fact that the largest eigenvalue of A is the maximum of $x'Ax$ for $x'x=1$ is basic in many proposed methods of finding the eigenvalues. From the standpoint of the algebraist, this use of the order properties of the real field to find (a rational approximation to) the roots of a polynomial with necessarily rational coefficients is suspect.

Against this it may be argued that to disregard the matric origin of the roots (for example, by direct computation of the coefficients of the characteristic equation in terms of determinants) is known to be hopelessly inefficient. Moreover, one wants the roots of the characteristic equation, and this is quite definitely a "nonrational" problem for which iterative methods of finding successive approximations are highly appropriate.

We are therefore led to seek a compromise in which "nearly rational" methods are used to reduce the complexity of the data presented to the computer (specifically, from $\frac{1}{2}n(n+1)$ numbers to $2n-1$) and then to accept the "irrationality" of the problem by solving for the roots and eigenvectors by iteration.

A purely mathematical result is now of interest[2]: Every symmetric matrix of rank r with elements in a principal ideal ring R is congruent with a matrix of the form (1), see above, where $\alpha_i=\beta_{i-1}=0$ if $i>r$. One notes that there is here, as elsewhere in this memorandum, no requirement that the matrix be definite. The congruence referred to is not restricted by a requirement of orthogonality, but the

[2] C. C. MacDuffee, Theory of matrices, p. 54 (Berlin, 1933).

reduced form (1) is highly rational (note the requirement on R) and is so easy to obtain that MacDuffee refers to a similar result for skewsymmetric matrices (given by Kronecker in 1883) to indicate a proof of the theorem. It is a little surprising, but definitely of interest, that requiring the transforming matrix (T in $T'AT=S$) to be orthogonal (a) does not prevent the existence of T (indeed, the diagonal form is a special case of (1)), (b) introduces no irrationality beyond a sequence of square roots, and (c) permits the step-by-step computation of T as a product of rotations in the coordinate planes. Factoring T into a product of more elementary transformations is quite analogous to the factorization of P and Q into elementary matrices, which is the key to the equivalence reduction $M \rightarrow PMQ$ of a general matrix M used in the successive elimination method of solving a system of linear equations.

17.4. Reduction to Triple Diagonal Form

The reduced form given in (1) can certainly be obtained when $n=1$ or 2, establishing satisfactorily the basis for the following induction. Nevertheless, it may be instructive to consider the geometry of the case $n=3$. To get the 1, 3 (and 3, 1) element of a symmetric matrix to be zero, we require that the (orthogonal) unit basis vectors f_1 and f_3 be conjugate: $f_1'Sf_3=s_{13}=0$. Starting from an orthogonal basis e_1, e_2, e_3 $(e_i'e_j=\delta_{ij})$, it is enough to take $f_1=e_1$ and f_3 in the intersection of the planes $e_1'Ax=0$ and $e_1'x=0$, that is, f_3 is in both the plane conjugate to e_1 and the plane orthogonal to e_1. For these planes to coincide, $e_1'Ae_2=e_1'Ae_3=0$ or $a_{12}=a_{13}=0$, and e_1 is already an eigenvector with value a_{11}. In the general case, an inductive step would be rendered unnecessary if it should happen that the r-dimensional space spanned by the first r basis vectors chosen was conjugate to its $(n-r)$-dimensional orthogonal space.

To establish the induction, let $a_{ij}^{(1)}=a_{ji}^{(1)}$, $i,j=1, \ldots, n$ be the elements of $A^{(1)}$, and suppose $a_{1k}^{(1)}$ is the first nonzero element in the sequence $a_{12}^{(1)}, a_{13}^{(1)}, \ldots, a_{1n}^{(1)}$. Interchanging the 2d and kth columns and then the 2d and kth rows, produces a new symmetric matrix $A^{(2)}$ with $a_{12}^{(2)}\neq 0$. One could also require $a_{12}^{(2)}=\max\{a_{12}^{(1)}, a_{13}^{(1)}, \ldots, a_{1n}^{(1)}\}$ if this should prove desirable to improve error estimates. Also there is no absolute need to actually change the position of the components as recorded in the machine, the specification of "next number in the column (row)" in the order code needs only to be adjusted. Also, if $a_{12}^{(1)}=0$, one later rotation that would be required in the general case will be unnecessary, so there is no need to "count" this operation.

If we now set $A^{(3)}=T_{23}'A^{(2)}T_{23}$, where

$$T_{23}=\left|\begin{array}{cccc} 1 & 0 & 0 & \\ 0 & c & -s & 0 \\ 0 & s & c & \\ & 0 & & 1_{n-3} \end{array}\right| \tag{3}$$

we get the following matrix, where, for convenience, we have omitted the superscript (2) on the elements of $A^{(2)}$.

$$A^{(3)}=\left|\begin{array}{cccccc} a_{11} & ca_{12}+sa_{13} & ca_{13}-sa_{12} & a_{14} & \cdots & a_{1n} \\ ca_{12}+sa_{13} & c(ca_{22}+sa_{23})+s(ca_{23}+sa_{33}) & c(ca_{23}-sa_{22})+s(ca_{33}-sa_{23}) & ca_{24}+sa_{34} & \cdots & ca_{2n}+sa_{3n} \\ ca_{13}-sa_{12} & c(ca_{23}+sa_{33})-s(ca_{22}+sa_{23}) & c(ca_{33}-sa_{23})-s(ca_{23}-sa_{22}) & ca_{34}-sa_{24} & \cdots & ca_{3n}-sa_{2n} \\ a_{14} & ca_{24}+sa_{34} & ca_{34}-sa_{24} & a_{44} & \cdots & a_{4n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{1n} & ca_{2n}+sa_{3n} & ca_{3n}-sa_{2n} & a_{4n} & \cdots & a_{nn} \end{array}\right|$$

Rule: To form $B=T'_{ij}AT_{ij}$, make the following replacements, where $(\ )_i$ is the ith column and $(\ )^i$ is the ith row: $(A)_i \rightarrow c(A)_i+s(A)_j$; $(A)_j \rightarrow c(A)_j-s(A)_i$; $(AT_{ij})^{(i)} \rightarrow c(AT_{ij})^i+s(AT_{ij})^j$; $(AT_{ij})^j \rightarrow c(AT_{ij})^j-s(AT_{ij})^i$. Use $c=[1+a_{i-1,j}(a_{i-1,i})^{-1}]^{-\frac{1}{2}}$ and $s=ca_{i-1,j}(a_{i-1,i})^{-1}$.

Taking

$$c=\frac{a_{12}}{\sqrt{(a_{12})^2+(a_{13})^2}} \quad \text{and} \quad s=\frac{a_{13}}{\sqrt{(a_{12})^2+(a_{13})^2}}, \tag{4}$$

or

$$c=[1+k^2]^{-\frac{1}{2}}, \quad s=ck \quad \text{and} \quad k=a_{13}(a_{12})^{-1}, \tag{5}$$

we have

$$a_{13}^{(3)}=0. \tag{6}$$

Evidently a rotation in the x_2-x_4 plane will make $a_{14}^{(4)}=0$ and retain the condition $a_{13}^{(4)}=0$ (since x_1 and x_3 are unaffected). After at most $n-2$ such rotations, $A^{(n)}$ will have the form

$$A^{(n)}=\left|\begin{array}{cccccc} * & * & 0 & 0 & \dots & 0 \\ * & & & & & \\ 0 & & & & & \\ 0 & & & & & \\ \cdot & & & B & & \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 & & & & & \end{array}\right| \tag{7}$$

where B is symmetric of order $n-1$. We now make the inductive hypothesis that an orthogonal congruence affecting only the rows and columns numbered 3, 4, . . ., n, will carry the matrix B to the reduced form (1). These do not disturb the zeros in the first row or column of $A^{(n)}$. (Note that the second row and column of $A^{(n)}$, which contain the first of B, are not "rotated".) We therefore have

Theorem 1. *A well-defined sequence of orthogonal congruences, affecting at each step only two rows and columns, will carry an arbitrary real symmetric matrix into a matrix S such that $s_{ij}=0$ if $|i-j|>1$.*

17.5 The Characteristic Equation

Forming $\lambda I-S$ for the reduced form (1) of S and expanding its determinant in terms of the last row and column, we get

$$f_i(\lambda)=(\lambda-\alpha_i)f_{i-1}(\lambda)-(\beta_{i-1})^2f_{i-2}(\lambda), \tag{8}$$

where $f_i(\lambda)$ is the determinant of the minor of order i in the upper-left corner of $\lambda I-S$. If we set

$$f_{-1}=0 \quad \text{and} \quad f_0=1, \tag{9}$$

the formula is valid for $i=1$ and 2.

If $\beta_{i-1}=0$, each of the $i-1$ real roots of $f_{i-1}(\lambda)=0$ is also a root of $f_i(\lambda)=0$ and hence of all $f_j(\lambda)=0$ for $j=i+1$, . . ., n. As noted in section 17.2, $\beta_i=0$ causes S to be the direct sum of a matrix of order i and one of order $n-i$, and so the problem reduces to two problems involving matrices of lower order. We therefore assume

$$\beta_i \neq 0 \quad \text{for } i=1, 2, \ . \ . \ ., n-1. \tag{10}$$

Under this assumption, for any chosen value of λ, no two consecutive $f_i(\lambda)$ can equal zero, since this would imply $f_0=0$. Hence $\lambda I-S$ is regularly arranged and the classical result of Darboux [3] on the signature of a quadratic form can be applied. This gives the second sentence of the following theorem.

Theorem 2. The functions $f_n, f_{n-1}, \ldots, f_0=1$ defined by the recursion formula (2) are a Sturm sequence if $\beta_i\neq 0$ for $i=1, 2, \ldots, n-1$. The number of eigenvalues of S greater than a, provided $f_n(a)\neq 0$, is the number of variations of sign in the sequence $f_n(a), f_{n-1}(a), \ldots, f_1(a), 1$. The roots of each $f_i(\lambda)=0$ are distinct and are separated by those of $f_{i-1}(\lambda)=0$.

To complete the proof of the theorem, we use induction on the following assertion, the case $i=3$ being easily established:

the $i-2$ roots of $f_{i-2}(\lambda)=0$ are distinct and properly separate the $i-1$ distinct roots of $f_{i-1}(\lambda)=0$. (11)

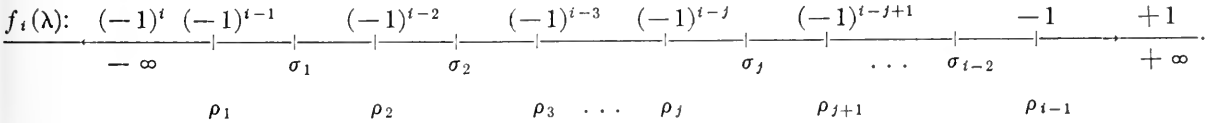
Explicitly, no root of $f_{i-2}(\lambda)=0$ is also a root of $f_{i-1}(\lambda)=0$. Let the roots of $f_{i-2}(\lambda)=0$ be $\sigma_1<\sigma_2<\ldots<\sigma_{i-2}$, and those of $f_{i-1}(\lambda)=0$ be $\rho_1<\rho_2<\ldots<\rho_{i-1}$, and we have assumed

$$\rho_1<\sigma_1<\rho_2<\sigma_2<\ldots<\sigma_{j-1}<\rho_j<\sigma_j<\rho_{j+1}<\ldots<\rho_{i-1}. \tag{12}$$

Writing the recursion formula in the form

$$f_i(\lambda)=(\lambda-\alpha_i)(\lambda-\rho_1)(\lambda-\rho_2)\ldots(\lambda-\rho_{i-1})-(\beta_i)^2(\lambda-\sigma_1)(\lambda-\sigma_2)\ldots(\lambda-\sigma_{i-2}), \tag{13}$$

we find on substituting ρ_j for λ that the signs of $f_i(\rho_j)$ are as indicated:



(14)

Since $f_i(\lambda)=0$ has exactly i real roots, they are distinct and are properly separated by the roots of $f_{i-1}=0$

All necessary properties that the sequence $f_n(\lambda), f_{n-1}(\lambda), \ldots, f_0=1$ be a Sturm sequence [4] are now evident.

17.6. Number of Multiplications Involved

The elements in the second and third columns of the matrix $A^{(3)}$ displayed in section 17.4 can be calculated with $4n+8$ multiplications. Since the (2,3) and (3,2) elements are equal and the sum of the (2,2) and (3,3) elements is unchanged by the rotation (so that the (3,3) element can be computed by addition and subtraction), this can be reduced to $4n+4$. Since $n-2$ of these rotations may be required at this step of the induction, we may need

$$\sum_{k=n}^{3}(k-2)(4k+4)=\frac{4}{3}(n^3-7n+6)$$

multiplications to reduce the original matrix to the form (1). The computation of $\frac{1}{2}(n-1)(n-2)$ values of c and s are also required. This could be done with $(n-1)(n-2)$ divisions, the same number of multiplications and half as many square roots. Neglecting powers of n below the third, we conclude that approximately $\frac{4}{3}n^3$ multiplications are required. Thus at a cost of about one and one-third matrix multiplications, the reduction to the triple diagonal form (1) can be accomplished.

Equations (2) involve only the squares of the β_i, and we suppose these have been computed and recorded. Then a single computation of the Sturm sequence requires $2(n-1)$ multiplications. If the eigenvalues are known to lie between -1 and $+1$, the largest eigenvalue can be found to s binary place accuracy (between ρ and $\rho+2^{-s}$) by $2(n-1)(s+1)$ multiplications. Using the method in its crudest form, one would require only $2(s+1)(n-1)n$ multiplications to get all n eigenvalues to this accuracy.

[3] See p. 57–58 of reference given in footnote 2.

[4] I am indebted to my colleague Walter Snyder for suggesting that these functions probably formed a Sturm sequence.

Because this is of the order of magnitude of a constant multiple of n^2, and hence comparable with terms previously neglected when $n \gg s$, we do not attempt a more careful estimate. For smaller values of n, say $n=s=40$, the crude determination of the eigenvalues could involve $2n^3$ multiplications or two matrix multiplications.

17.7. Relation to Other Methods

An inspection of other methods proposed for the eigenvalue problem will show that in a number of cases the triple diagonal form and the recursion formulas for the principal minors both occur.[5] Here the fundamental paper of C. Lanczos deserves mention.[6] [7] (This also contains much material relating the question to differential and integral as well as algebraic problems.) The same two basic devices are also to be found in a recently published paper by W. Karush [8] and in the work of Hestenes and Stiefel, as presented elsewhere in this volume.

[5] Indeed for operators in Hilbert space the triple diagonal form is well known under the name "Jacobi Form."

[6] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Research NBS **45**, 255–282 (1950) RP 2133.

[7] Entirely through oversight, this reference was omitted when the material was presented to the Symposium. This was in spite of the fact that, although I had not seen the paper previously, it had been called to my attention during the conference. Dr. W. Karush kindly rectified my error in the discussion.

[8] W. Karush, An iterative method for finding characteristic vectors of a symmetric matrix, Pacific J. Math. **1**, 233–248 (1951).

18. Computations Relating to Inverse Matrices

Jack Sherman[1]

Introduction

In many practical problems it is frequently desired to obtain inverses of matrices that differ from one another in some systematic manner, for example, in the elements of a particular row or column. It would obviously be advantageous to develop computational methods for obtaining these inverses directly from one another, rather than from the original matrix. Some pertinent results are described below.

Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix[2]

A simple relationship has been derived by means of which the inverse of a given matrix can be "adjusted," as a result of changing one element in the original matrix. The set of equations may be written as

$$B_{rj}=b_{rj}-\frac{b_{rR}b_{Sj}\Delta a_{RS}}{1+b_{SR}\Delta a_{RS}}, \qquad r=1,2,\ldots,n, \quad j=1,2,\ldots,n,$$

provided that $1+b_{SR}\Delta a_{RS}\neq 0$.

In these equations B_{rj} denote the elements of (B), the inverse of (A); b_{rj} denote the elements of (b), the inverse of (a). Δa_{RS} denotes the change in the element a_{RS}, that is $A_{RS}=a_{RS}+\Delta a_{RS}$. It is seen from the preceding equations that if $\Delta a_{RS}=-1/b_{SR}$, then (A) becomes singular.

Adjustment of an Inverse Matrix Corresponding to Changes in the Elements of a Given Column[3]

If (A) differs from (a) in the elements of the kth column, then the equations for computing the elements of (B) from (b) are as follows:

$$\begin{aligned} B_{ij}&=b_{ij}-z_iB_{kj}, && i=1,2,\ldots,k-1,k+1.\ldots,N, \quad j=1,2,\ldots,N, \\ B_{kj}&=b_{kj}/z_k, && j=1,2,\ldots,N, \end{aligned}$$

in which

$$z_i=\sum_{r=1}^{N} b_{ir}A_{rk}, \qquad i=1,2,\ldots N.$$

The foregoing equations can be conveniently utilized to carry out the numerical computations on IBM machines for obtaining (B) from (b) and the elements A_{rk}.

Equations analogous to the above can easily be written for the case that (A) differs from (a) in the elements of a given row rather than a given column.

[1] The Texas Co., Beacon, N. Y.

[2] Abstracted from an article by Jack Sherman and Winifred J. Morrison, The Annals of Mathematical Statistics, 21 (March 1950).

[3] Presented at the Twelfth Summer Meeting of the Institute of Mathematical Statistics, Boulder, Colo., 1949.

Computation of an Inverse Matrix by Means of IBM Machines

The foregoing relationship for obtaining the inverse of a matrix differing from a given matrix only in the elements of a given column can be utilized to provide a convenient means of computing an inverse matrix.

Starting with a given matrix (a) and its inverse (b), the inverse (β) of any other matrix (α) may be obtained by N applications of the foregoing procedure for replacing the elements of a given column. In other words, (a) may be transformed into (α) by N steps, the first involving the replacement of the elements of the first column of (a) by the elements of the first column of (α), the second step involving the replacement of the elements of the second column of (a) by those in the second column of (α), etc. Corresponding to each of these changes the inverse may be computed according to the foregoing methods.

If (a) and (b) are taken to be the unit matrices, then the method described of systematically replacing the column of the unit matrix by those of (α) is exactly equivalent to the familiar method of systematic elimination. However, if the matrix (α) to be inverted is similar to a matrix (a) that has previously been inverted, then better control of the computations can be achieved by using the presently described method.

The method of obtaining an inverse by the procedure described in this paper is being carried out on IBM machines at the present time in connection with spectrometric analyses. It has been found to be more efficient than the orthodox methods of systematic elimination. Furthermore, this method provides considerable flexibility, and obvious saving of computations in cases where (α) differs from (a) only in a few of the N columns.

If matrix (a) is transformed by deleting a row and column or by augmenting a row and column, the corresponding inverses of orders $N-1$ and $N+1$, respectively, can be easily obtained from (b). In the case of a symmetrical matrix, it is possible to obtain simple relationships for computing (β) from (b) for the case that (α) differs from (a) in the elements of the kth row and column.

19. Results of Recent Experiments in the Analysis of Periods Carried Out in the Istituto Nazionale per le Applicazioni del Calcolo[1]

Galtano Fichera[2]

Let $f(t)$ be a function that has empirically given values, real or complex in the interval $-1 \leq t \leq +1$. The analysis of the periods of the function $f(t)$ means the following: Evaluation of the first natural number n and the constants $\gamma_1, \gamma_2, \ldots, \gamma_n; z_1, z_2, \ldots, z_n$, such that

$$\left| f(t) - \sum_{k=1}^{n} \gamma_k e^{z_k t} \right| < \epsilon, \tag{1}$$

where ϵ is a known positive constant.

A method of the analysis of the periods will be outlined briefly. As $f(t)$ is given empirically, it is not admissible to use expressions that involve differential quotients of $f(t)$, but definite integrals of $f(t)$ or products of $f(t)$ with other known functions may be used. The number n must not be too large. The method is the following. If we had exactly

$$f(t) = \sum_{k=1}^{n} \gamma_k e^{z_k t},$$

then $f(t)$ would represent a solution of a differential equation with constant coefficients

$$\sum_{k=0}^{n} C_k f^{(k)}(t) = 0, \quad C_n = 1. \tag{2}$$

If $\varphi_s(t)$, $(s=1,2,3, \ldots)$ represents a complete system of functions in the interval $[-1,+1]$ in the sense of Hilbert, then (2) is equivalent to

$$\int_{-1}^{1} \varphi_s(t) \sum_{k=0}^{n} C_k f^{(k)}(t) dt = 0, \quad s = 1, 2, 3, \ldots,$$

which, after a number of partial integrations, can be written as

$$\sum_{i=1}^{n} \varphi_s^{(i-1)}(1) \beta_i + \sum_{i=1}^{n} \varphi_s^{(i-1)}(1) \alpha_i + \sum_{k=0}^{n-1} (-1)^k C_k \int_{-1}^{1} \varphi_s^{(k)}(t) f(t) dt = (-1)^{n-1} \int_{-1}^{1} \varphi_s^{(n)}(t) f(t) dt, \quad s = 1, 2, 3, \ldots, \tag{3}$$

with

$$\beta_i = (-1)^{i-1} \sum_{k=i}^{n} C_k f^{(k-i)}(1), \quad \alpha_i = (-1)^i \sum_{k=i}^{n} C_k f^{(k-i)}(-1).$$

But, as $f(t)$ is only given empirically, the integration in (3) can be performed, but like the α_i and β_i cannot be evaluated as they involve differential quotients of $f(t)$ at $t = \pm 1$. Therefore, one considers the α_i and β_i in (3) as new unknown parameters. (3) is then a system of an infinite number of linear equations with $3n$ unknown parameters α_i, β_i, $(i=1,2, \ldots n)$ and C_k $(k=0,1, \ldots n-1)$. The matrix of the coefficients of the system (3) has the rank $3n$, if, for instance, $\varphi_s(t) = e^{\lambda_s t}$ and the sequence of the λ_s has the property

$$\operatorname{Max} \lim_{s \to \infty} \lambda_s = +\infty, \quad \operatorname{Min} \lim_{s \to \infty} \lambda_s = -\infty. \tag{4}$$

[1] This is a report on an address delivered by M. Picone and F. Conforto at a meeting of the G. A. M. M. at Freiburg i. Br., March 1951.

[2] University of Trieste and Italian Institute for the Application of Calculus.

Provided, that $f(t)$ does not satisfy a linear homogeneous differential equation with constant coefficients of the order $m<n$, the sequence λ_s can be, for instance, the sequence $0, -1, +1, -2, +2, -3, \ldots$ In other words, one considers for each empirically given $f(t)$ the system (3) with $\varphi_s(t)=e^{\lambda_s t}$, where (4) is valid for the λ_s.

Numerous numerical experiments that have been carried out in the Istituto Nazionale per le Applicazioni del Calcolo have shown that the system (3) for a number of p equations gives solutions, that tend to the exact values of the C_k with increasing p.

If, therefore, $f(t)$ is given empirically and the C_k $(k=0,1,2, \ldots)$ are unknown, then it is indicated to solve the system of p equations of the system (3) by the method of least squares for $n=1,2, \ldots$ If n is the first number, such that for increasing p the c_k converge to C_k, then it can be conjectured that $f(t)$ approximates an integral of (2), formed with these coefficients C_k. The roots z_k of the characteristic equation of this differential equation can be determined then and the γ_k are solutions of the system of p equations

$$\int_{-1}^{1} e^{\lambda_s t} f(t)\, dt = \sum_{k=1}^{n} \gamma_k \int_{-1}^{1} e^{\lambda_s t} e^{z_k t}\, dt, \quad s=1, 2, 3, \ldots$$

This system can be solved by the method of least squares, provided that for increasing p the solution tends to a limit γ_k. If the γ_k and z_k are evaluated, one can check if (1) is satisfied. In case it is satisfied, the problem can be considered as solved, otherwise, one has to perform the same procedure with $n+1$, $n+2, \ldots$ In case n is too large, the problem has to be considered as unsolvable.

U. S. GOVERNMENT PRINTING OFFICE, 1953